AGENO SCHOOL OF BUSINESS

Golden Gate University

A Statistical Analysis on Super Store Sales Data

MSBA 320: Advanced Statistical Analysis with R & Python

Summer 2020

Buse Bastug #0598594

Final Project

Submitted to Professor Siamak Zadeh

Table of Contents

Introduction

This paper aims to conduct a wide variety of analysis about Superstore Sales Data which is between 2011 and 2015. Superstore is a fictitious company and the dataset has been especially created for data visualization practice. It is a very popular dataset to use in Tableau and MicroStrategy and it can be retrieved from Kaggle. The dataset lists 51290 entries in 24 columns and has 4 KPIs (Key Performance Indicator) such as Sales, Profit, Discount and Shipping Cost.

In this paper, descriptive statistics were used to summarize the data and have an overview of existing parameters, detect any possible outliers and generate visual plots. This paper aims sales department in mind and a time series analysis to see sales trends in each year with correlation and regression analysis conclusion.

Data Collection

Row ID Order ID Customer ID Product ID	Nominal. Assigned to each product and customer.
Customer Name Product Name	Nominal.
Segment Category Sub-Category	Non-numerical. 3 segment and 3 categories. 17 different sub-category names.
Order Date Ship Date	Numeric. Assigned to each order and ship time.
Sales Quantity Discount Profit Shipping Cost	Numeric. Have values for 4 years between 2011-2015
City State Country Market Region	Non-numeric. Have all values for each city and state with 7 market and 13 regions.

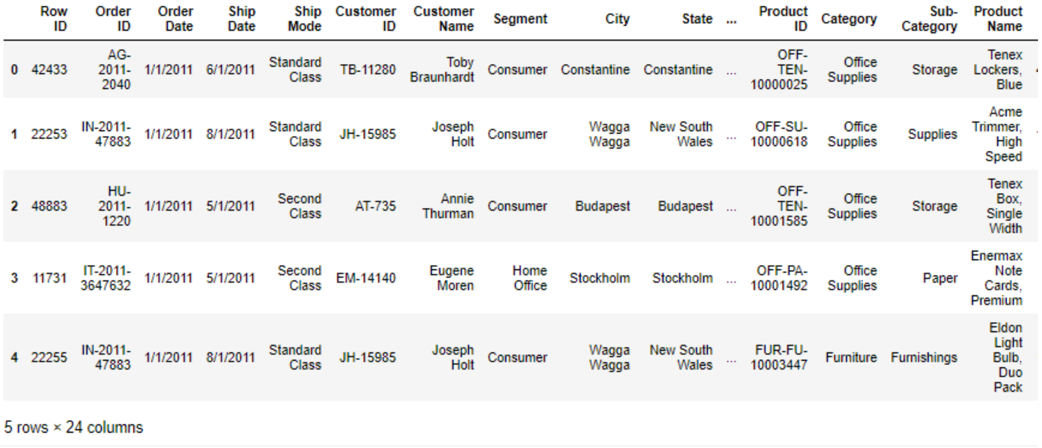
Order Priority Ship Mode Postal Code	Non-numerical. 4 priority shipping models and order priority. Postal code is assigned to each customer.

Descriptive Statistic

Descriptive statistic provides us a simple summary about our data. It uses data mining and data aggregation techniques for providing summary of the past actions by focusing on "what happened?" question. With this gathering and summarizing historical data focus, descriptive statistic can provide a better understanding of key business metrics and present the overall picture of the company in a understandable way to business executives or any users. Descriptive statistic is a significant and the first step in advanced statistic for any further action that a company would take based on data analysis.

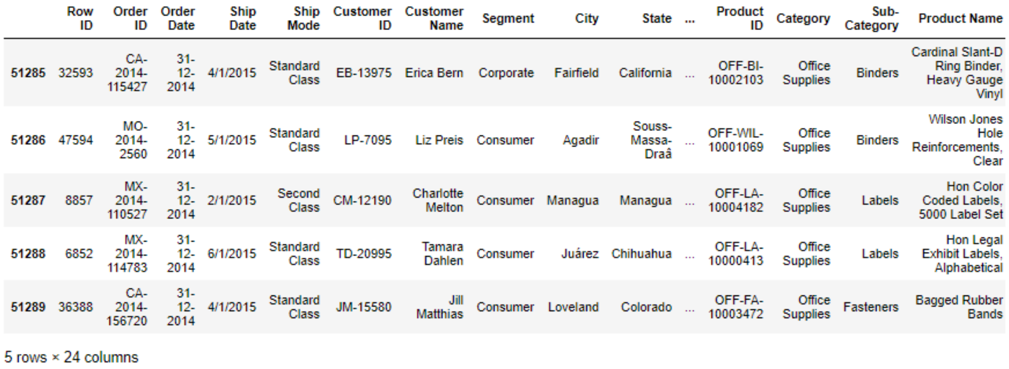
Data cleansing, and transformations provided in our data set as follows:

- Missing values were removed.
- Shipping Date and Order Date values were transformed from object to date time.



	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	City	State	...	Product ID	Category	Sub-Category	Product Name
0	42433	AG-2011-2040	1/1/2011	6/1/2011	Standard Class	TB-11280	Toby Braunhardt	Consumer	Constantine	Constantine	...	OFF-TEN-10000025	Office Supplies	Storage	Tenex Lockers, Blue
1	22253	IN-2011-47883	1/1/2011	8/1/2011	Standard Class	JH-15985	Joseph Holt	Consumer	Wagga Wagga	New South Wales	...	OFF-SU-10000618	Office Supplies	Supplies	Acme Trimmer, High Speed
2	48883	HU-2011-1220	1/1/2011	5/1/2011	Second Class	AT-735	Annie Thurman	Consumer	Budapest	Budapest	...	OFF-TEN-10001585	Office Supplies	Storage	Tenex Box, Single Width
3	11731	IT-2011-3647632	1/1/2011	5/1/2011	Second Class	EM-14140	Eugene Moren	Home Office	Stockholm	Stockholm	...	OFF-PA-10001492	Office Supplies	Paper	Enermax Note Cards, Premium
4	22255	IN-2011-47883	1/1/2011	8/1/2011	Standard Class	JH-15985	Joseph Holt	Consumer	Wagga Wagga	New South Wales	...	FUR-FU-10003447	Furniture	Furnishings	Eldon Light Bulb, Duo Pack

5 rows × 24 columns

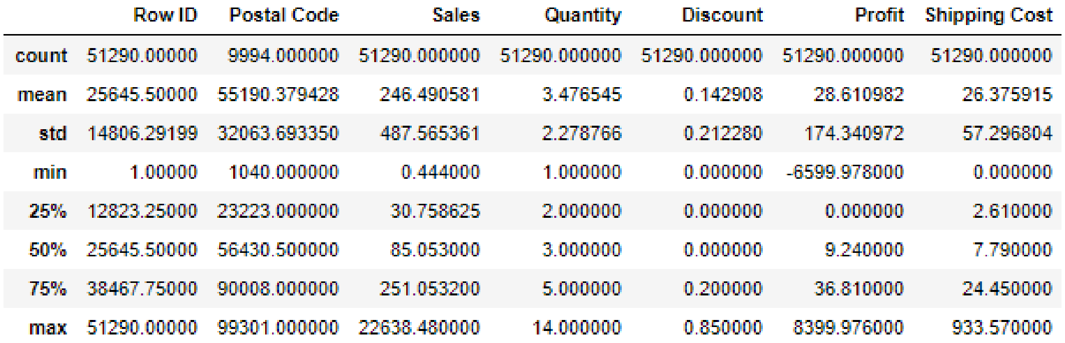


	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	City	State	...	Product ID	Category	Sub-Category	Product Name
51285	32593	CA-2014-115427	31-12-2014	4/1/2015	Standard Class	EB-13975	Erica Bern	Corporate	Fairfield	California	...	OFF-BI-10002103	Office Supplies	Binders	Cardinal Slant-D Ring Binder, Heavy Gauge Vinyl
51286	47594	MO-2014-2560	31-12-2014	5/1/2015	Standard Class	LP-7095	Liz Preis	Consumer	Agadir	Souss-Massa-Draâ	...	OFF-WIL-10001069	Office Supplies	Binders	Wilson Jones Hole Reinforcements, Clear
51287	8857	MX-2014-110527	31-12-2014	2/1/2015	Second Class	CM-12190	Charlotte Melton	Consumer	Managua	Managua	...	OFF-LA-10004182	Office Supplies	Labels	Hon Color Coded Labels, 5000 Label Set
51288	6852	MX-2014-114783	31-12-2014	6/1/2015	Standard Class	TD-20995	Tamara Dahlen	Consumer	Juárez	Chihuahua	...	OFF-LA-10000413	Office Supplies	Labels	Hon Legal Exhibit Labels, Alphabetical
51289	36388	CA-2014-156720	31-12-2014	4/1/2015	Standard Class	JM-15580	Jill Matthias	Consumer	Loveland	Colorado	...	OFF-FA-10003472	Office Supplies	Fasteners	Bagged Rubber Bands

5 rows × 24 columns

Table 1: Top and bottom 5 entries of Super Store Sales Data..

Table 2 provides the summary of our existing variables.



	Row ID	Postal Code	Sales	Quantity	Discount	Profit	Shipping Cost
count	51290.00000	9994.000000	51290.000000	51290.000000	51290.000000	51290.000000	51290.000000
mean	25645.50000	55190.379428	246.490581	3.476545	0.142908	28.610982	26.375915
std	14806.29199	32063.693350	487.565361	2.278766	0.212280	174.340972	57.296804
min	1.00000	1040.000000	0.444000	1.000000	0.000000	-6599.978000	0.000000
25%	12823.25000	23223.000000	30.758625	2.000000	0.000000	0.000000	2.610000
50%	25645.50000	56430.500000	85.053000	3.000000	0.000000	9.240000	7.790000
75%	38467.75000	90008.000000	251.053200	5.000000	0.200000	36.810000	24.450000
max	51290.00000	99301.000000	22638.480000	14.000000	0.850000	8399.976000	933.570000

Table 2: Summary Statistics

Product Level Analysis

Product level analysis is important for a company to see sales trend over segments, categories and sub-categories. It allows us to do a deeper analysis of customers purchase pattern and detect if there is any area to be improved as product category in order to be more profitable. Thus, the visualization below will help us to see existing patterns and overall picture of the company.

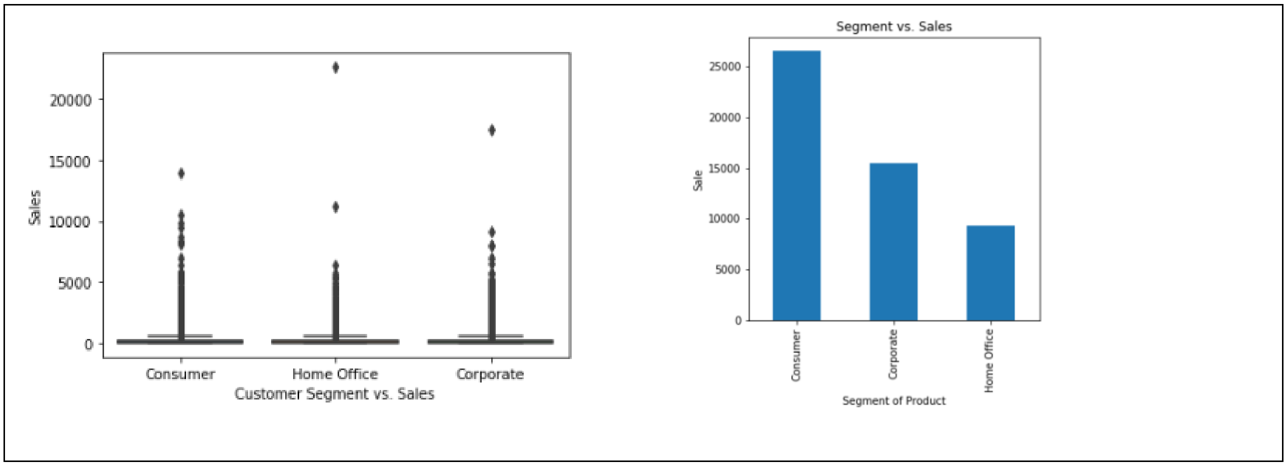


Figure1: Customer Segment vs. Sales

Figure 1 provides us that Consumer section is the first customer segment at the company. Checking the other parameters, office supplies section is the best seller and the best seller products are shown as sub-categories in Figure 2.

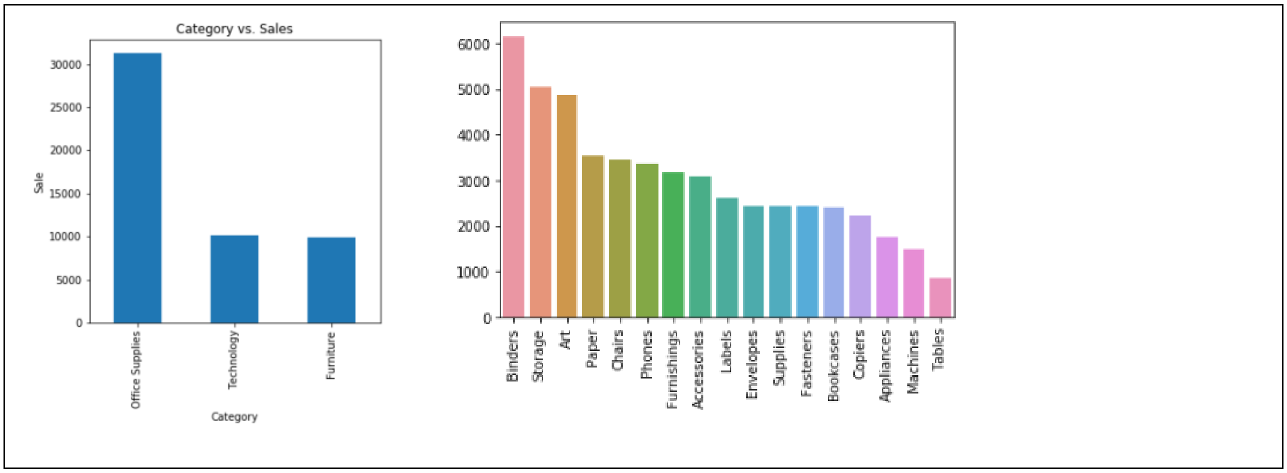


Figure2: Product Category vs. Sales

Order Priority and Ship Mode are important variables and checking their relationship with sales and profit KPIs, it provides us some interesting insights. Most of the company products are ordered and shipped as high priority. The existing outliers for medium and high order priority both in sales and profit boxplot gives us the clue that some of the products have been ordered from a very high price than usual as much as some of the products have been ordered from a very low price which affected profit in a negative way. This could be interpreted as the existence of loyal and regular customers.

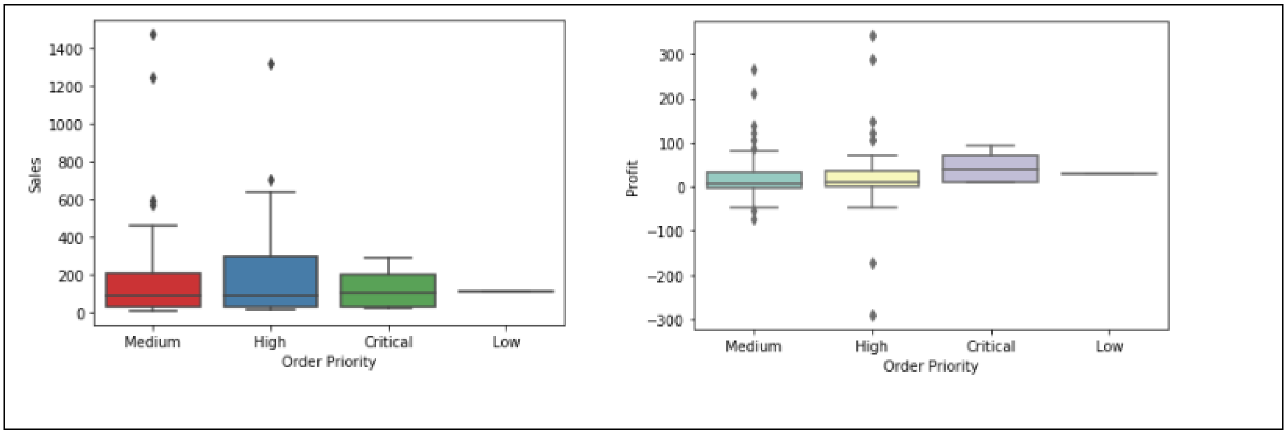


Figure3:Order Priority vs. Sales & Profit

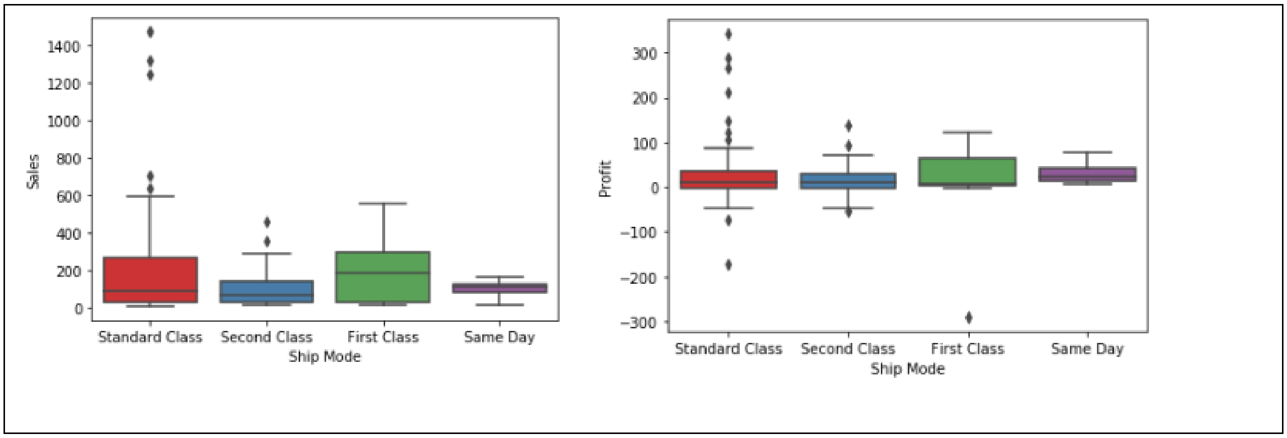


Figure4: Ship Mode vs. Sales & Profit

First class ship mode is the most profitable for the company. There is only one outlier which is quite interesting that can be seen in first class and profit boxplot that the order has been shipped as first class with a very low price. Standard class and first class are the most preferable methods at the company.

On the other hand, checking the category section by sales and profit variables, we see that technology category is the most profitable one however, it is not the first category in sales. Thus, the company needs to be more focused on increasing the sale of technology products. Some of the steps could be taken such as offering discounted prices or following a market strategy for sake of technology products.

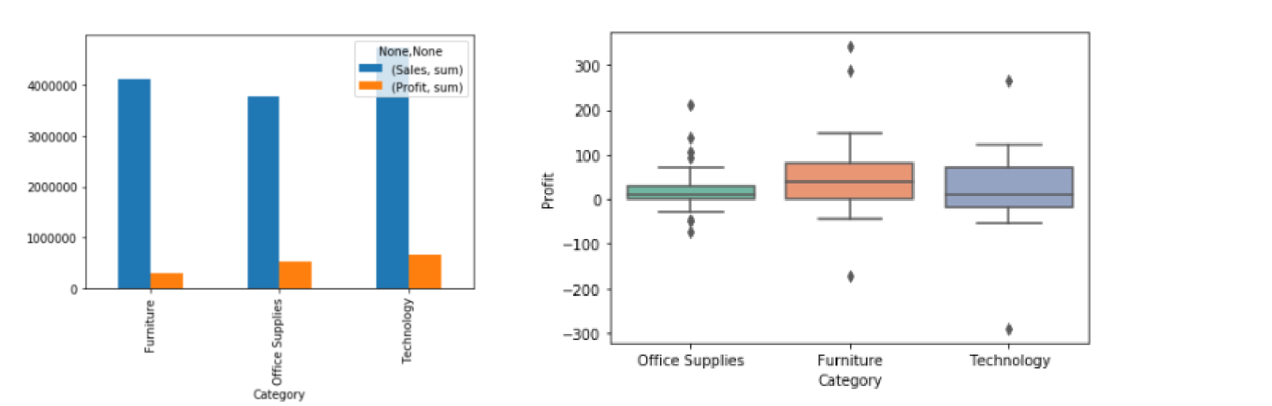


Figure5: Most Profitable Categories

Market Level Analysis

Super Store has 7 markets, 13 regions, country, state and city variables in its dataset. In order to understand each of these variable based on product level, we need to do a deeper analysis by selecting sales as our dependent variables and the other parameters as independent.

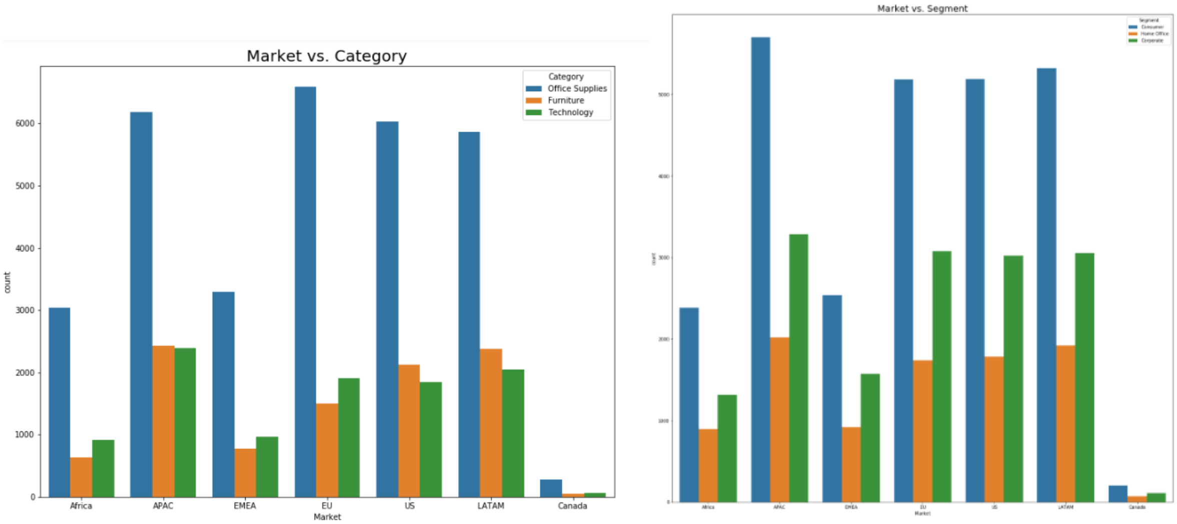
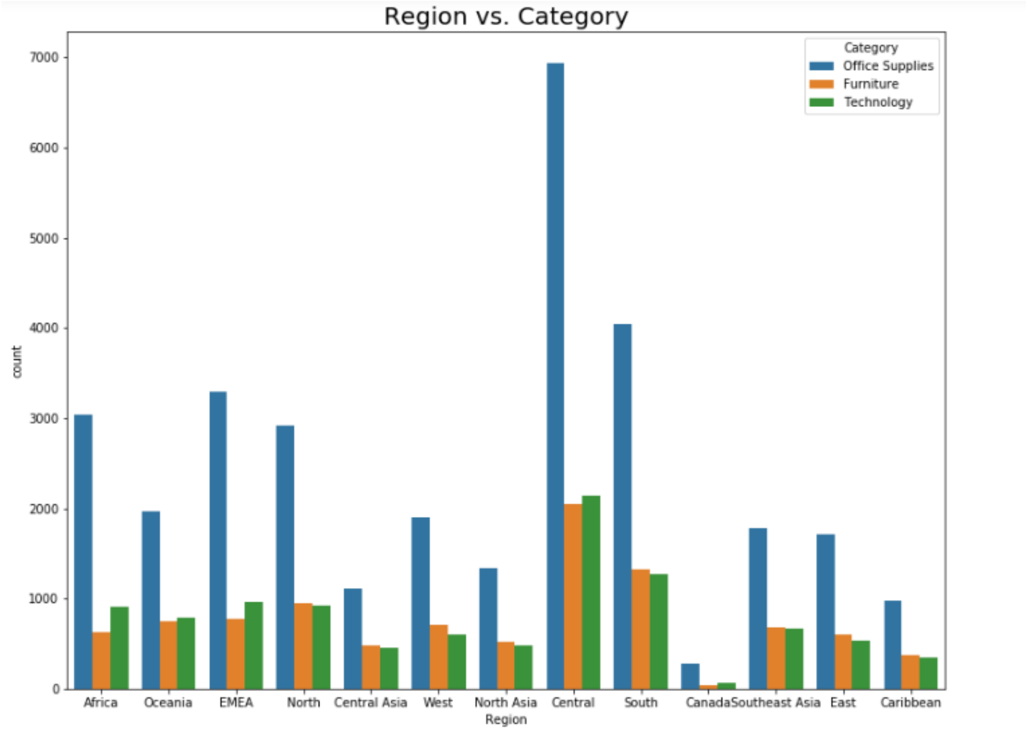


Figure6:Customer Segment and Product Category based on Market and Region



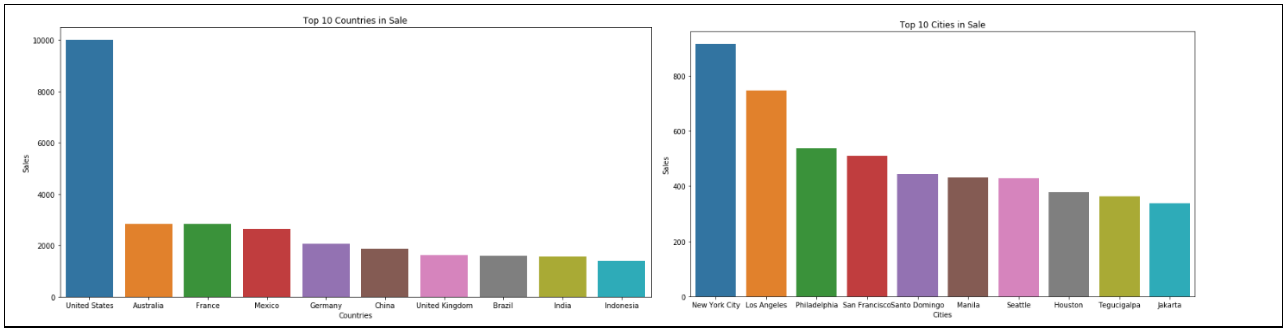


Figure7: Top 10 countries and cities in Sales

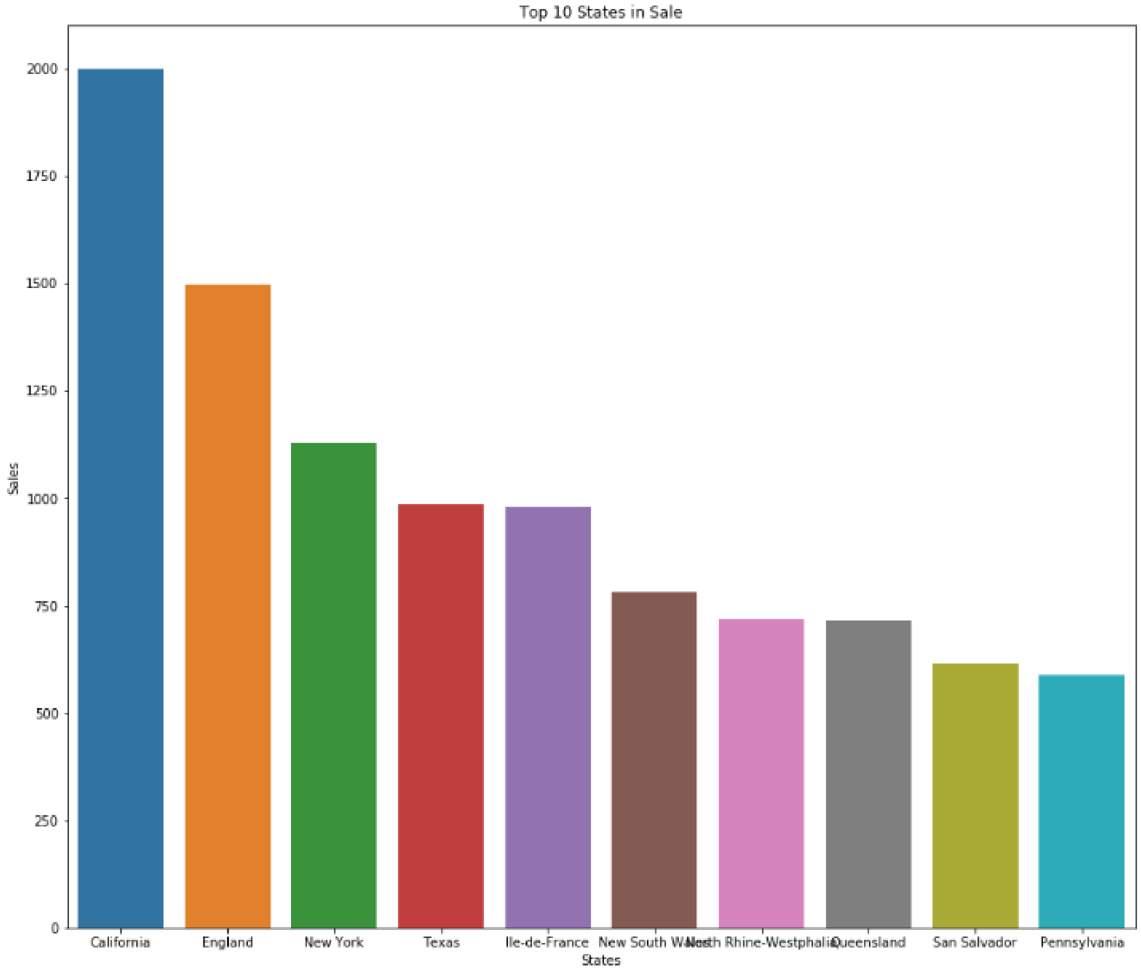


Figure8: Top 10 States in Sales

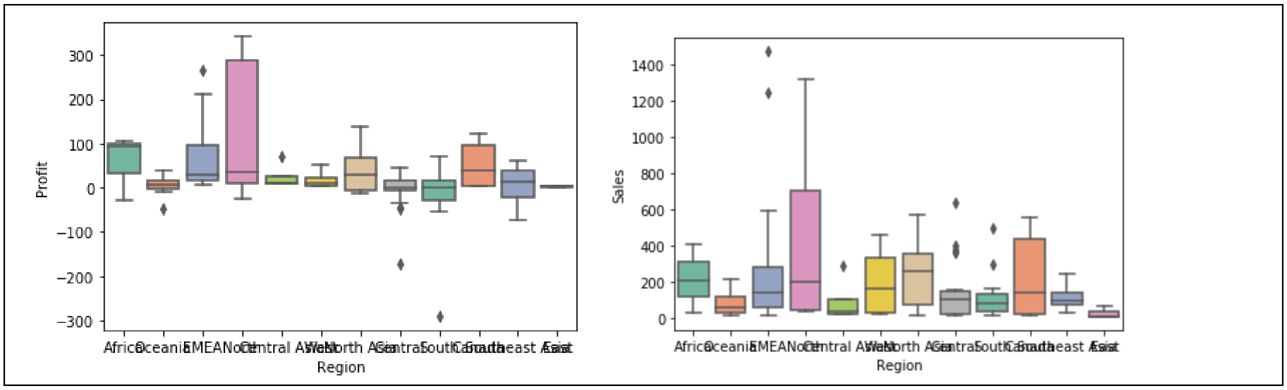


Figure9:Sales and Profit Among Regions

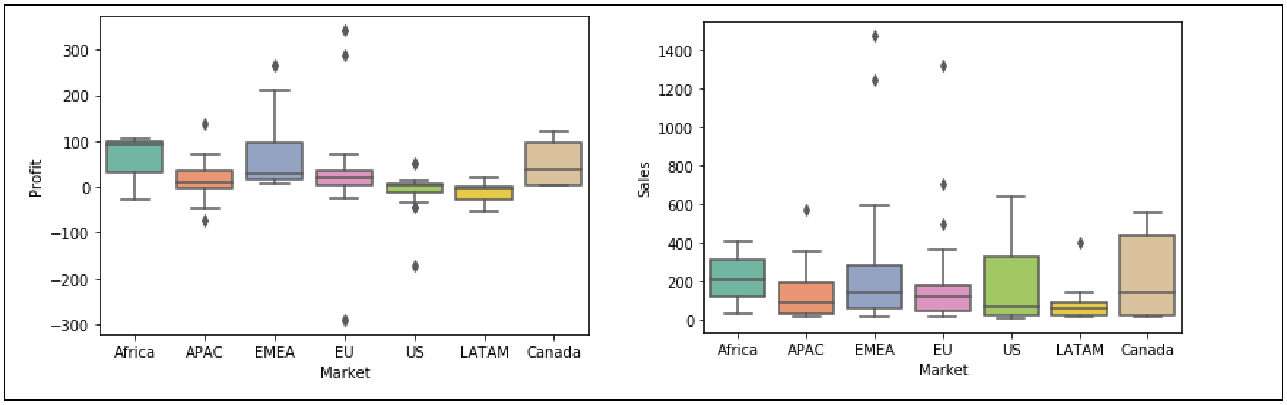


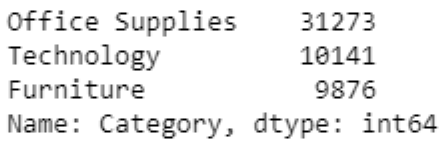
Figure10:Sales and Profit Among Market

North region and especially the U.S. market is at the top of sales. However, profit is significantly less in same region. The company needs to increase technology products sale in this market in order to increase its profitability.

Time Series Analysis

As we mentioned earlier, Super Store has three different categories:

- Furniture
- Office Supplies
- Technology



```
Office Supplies    31273
Technology         10141
Furniture           9876
Name: Category, dtype: int64
```

In order to get the best insights from our data, we ran the time series analysis for each category. In this way, we can also detect if there is any seasonality in different categories.

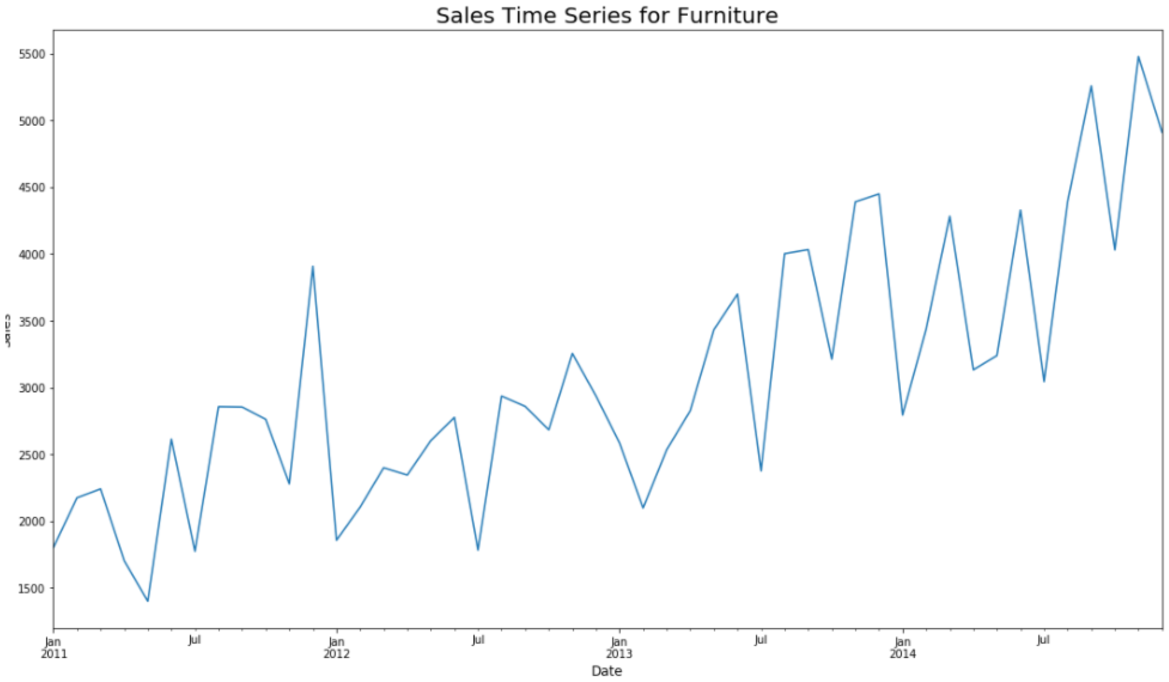


Figure11: Time Series Analysis for Furniture

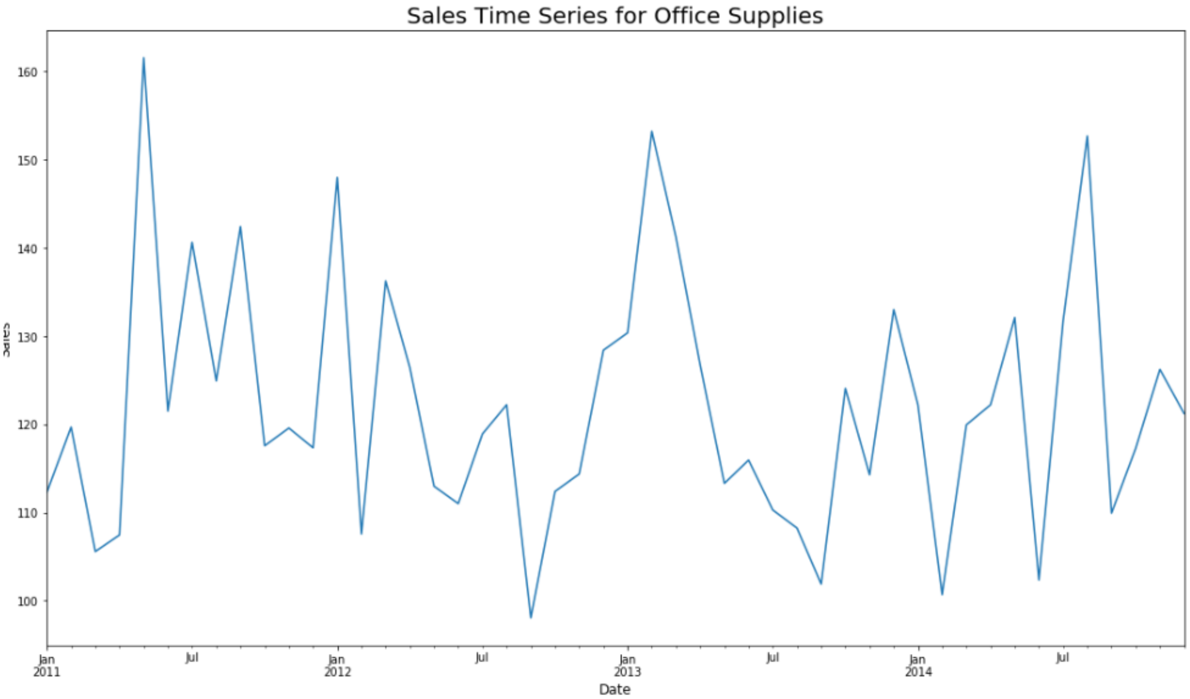


Figure12: Time Series Analysis for Office Supplies

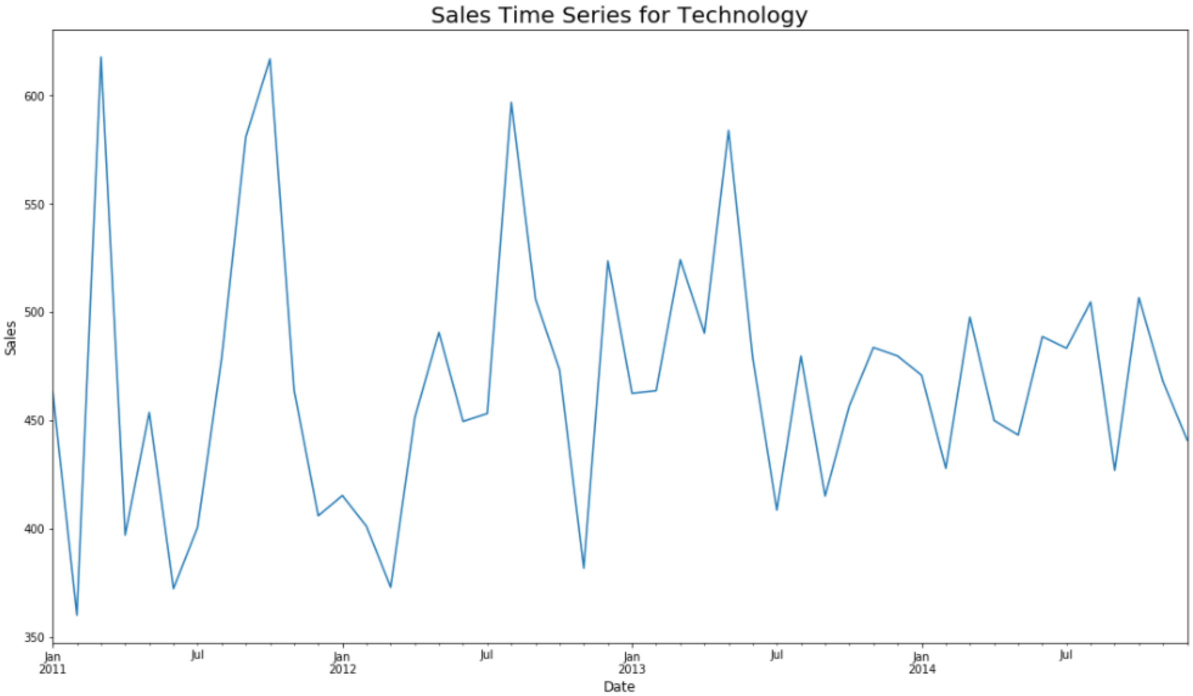


Figure13: Time Series Analysis for Technology

According to time series analysis, there is a seasonality in each category. Sales significantly decreases each year in January and then it starts to increase.

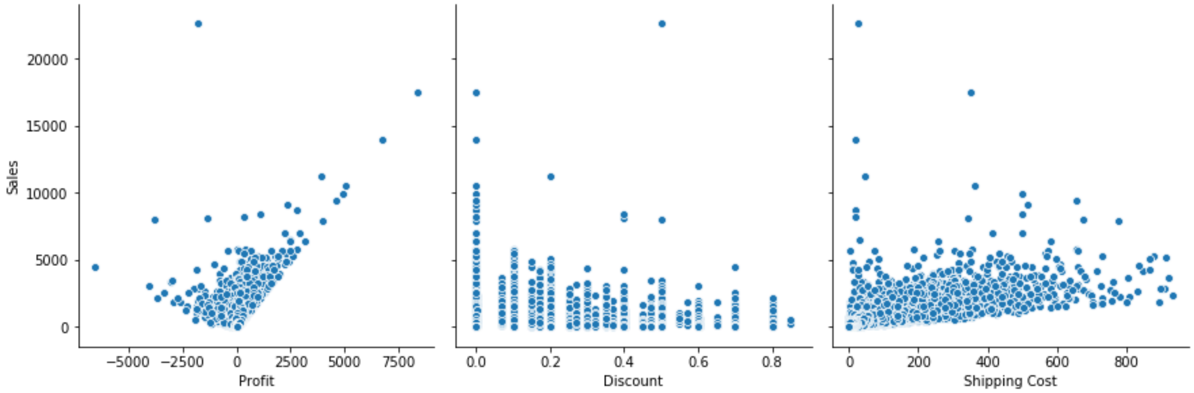
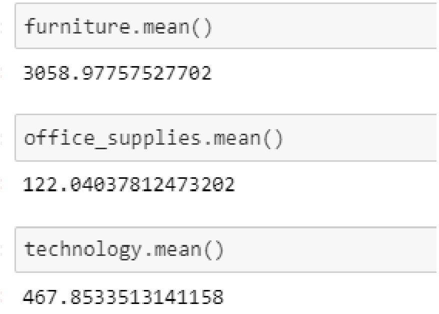


Figure12:Pairplots for Discount, Profit and Shipping Cost relationship with Sales



```
furniture.mean()
```

```
3058.97757527702
```

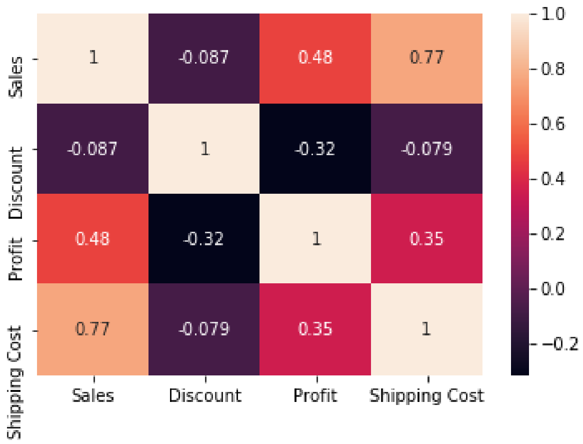
```
office_supplies.mean()
```

```
122.04037812473202
```

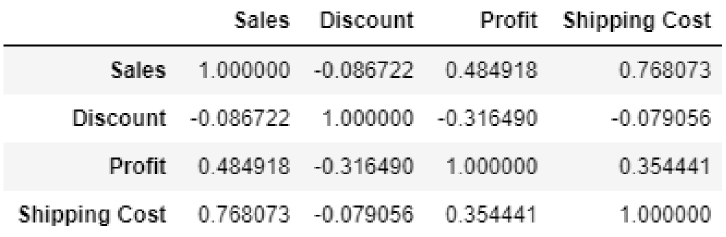
```
technology.mean()
```

```
467.8533513141158
```

Correlation Analysis

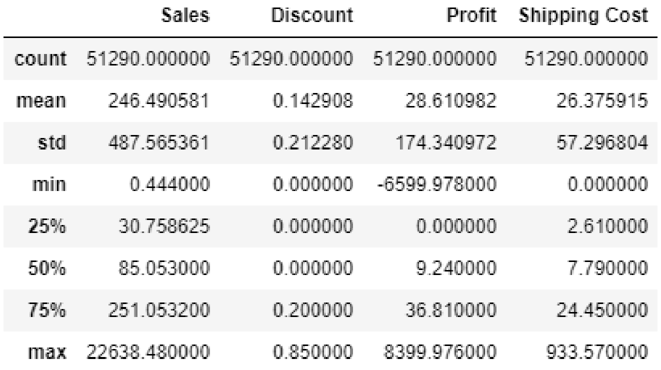


There is a significant correlation between Sales and Shipping Cost variables.



	Sales	Discount	Profit	Shipping Cost
Sales	1.000000	-0.086722	0.484918	0.768073
Discount	-0.086722	1.000000	-0.316490	-0.079056
Profit	0.484918	-0.316490	1.000000	0.354441
Shipping Cost	0.768073	-0.079056	0.354441	1.000000

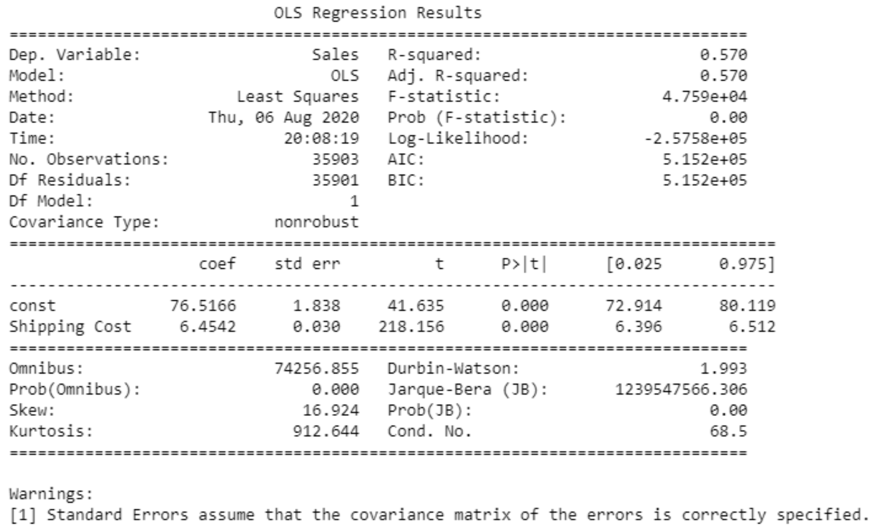
Table3: Correlation



	Sales	Discount	Profit	Shipping Cost
count	51290.000000	51290.000000	51290.000000	51290.000000
mean	246.490581	0.142908	28.610982	26.375915
std	487.565361	0.212280	174.340972	57.296804
min	0.444000	0.000000	-6599.978000	0.000000
25%	30.758625	0.000000	0.000000	2.610000
50%	85.053000	0.000000	9.240000	7.790000
75%	251.053200	0.200000	36.810000	24.450000
max	22638.480000	0.850000	8399.976000	933.570000

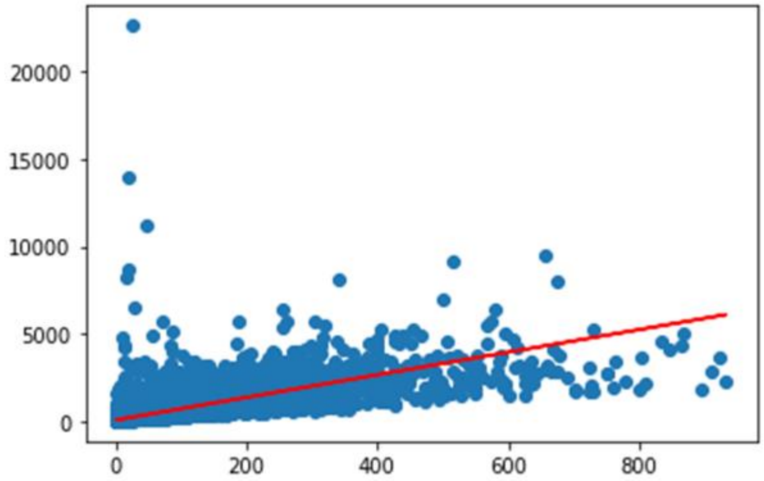
Table4: Summary Statistic of important variables

In order to understand this relationship better, we ran a simple linear regression:



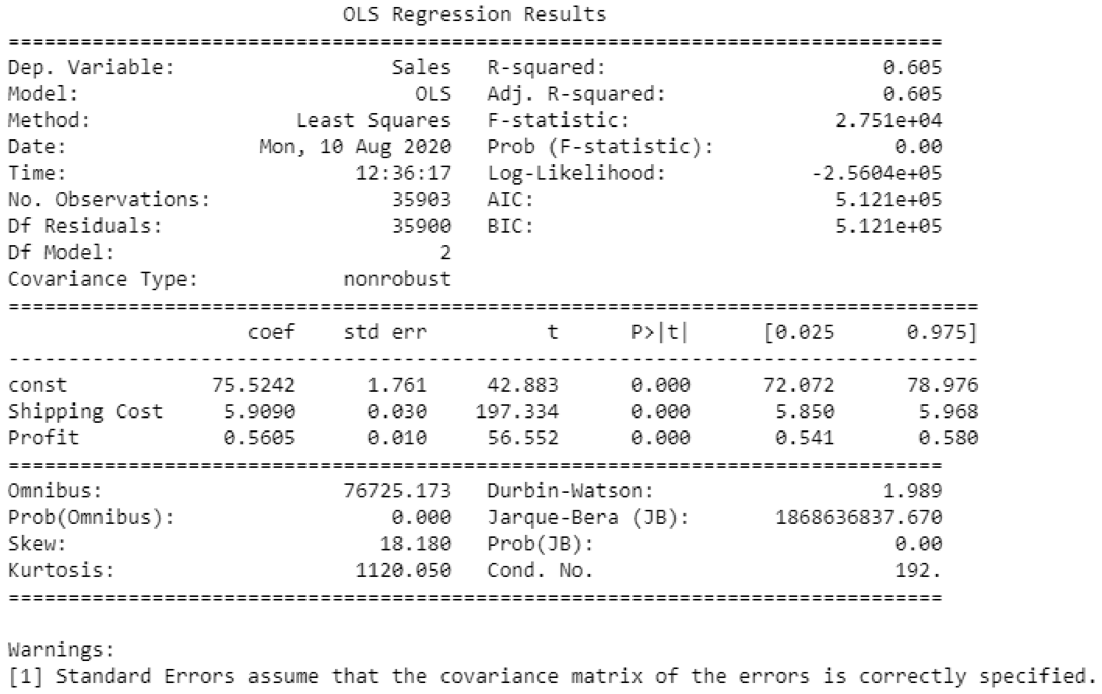
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.570
Model:                            OLS   Adj. R-squared:                  0.570
Method:                 Least Squares   F-statistic:                 4.759e+04
Date:                Thu, 06 Aug 2020   Prob (F-statistic):               0.00
Time:                        20:08:19   Log-Likelihood:            -2.5758e+05
No. Observations:               35903   AIC:                         5.152e+05
Df Residuals:                   35901   BIC:                         5.152e+05
Df Model:                           1
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const            76.5166      1.838     41.635      0.000      72.914      80.119
Shipping Cost     6.4542      0.030    218.156      0.000       6.396       6.512
==============================================================================
Omnibus:                    74256.855   Durbin-Watson:                   1.993
Prob(Omnibus):                  0.000   Jarque-Bera (JB):       1239547566.306
Skew:                          16.924   Prob(JB):                         0.00
Kurtosis:                     912.644   Cond. No.                         68.5
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```



R-Squared and Adjusted R-Squared are same value. This gives us the signal that our dependent and independent variable are relevant. Prob F-statistic is zero and F statistic is large so we can reject the null hypothesis and accept the alternative hypothesis. Thus, there is a linear relationship between Shipping Cost and Sales.

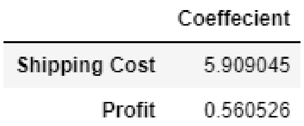
Predictive Analysis



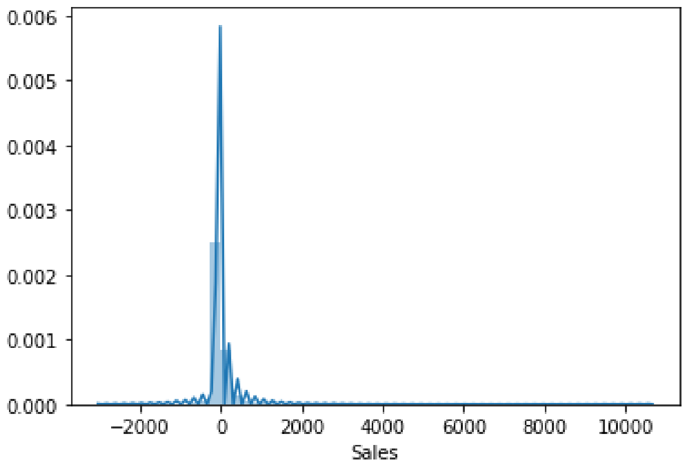
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.605
Model:                            OLS   Adj. R-squared:                  0.605
Method:                 Least Squares   F-statistic:                 2.751e+04
Date:                Mon, 10 Aug 2020   Prob (F-statistic):               0.00
Time:                        12:36:17   Log-Likelihood:            -2.5604e+05
No. Observations:               35903   AIC:                         5.121e+05
Df Residuals:                   35900   BIC:                         5.121e+05
Df Model:                           2
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const            75.5242      1.761     42.883      0.000      72.072      78.976
Shipping Cost     5.9090      0.030    197.334      0.000       5.850       5.968
Profit            0.5605      0.010     56.552      0.000       0.541       0.580
==============================================================================
Omnibus:                    76725.173   Durbin-Watson:                   1.989
Prob(Omnibus):                  0.000   Jarque-Bera (JB):      1868636837.670
Skew:                          18.180   Prob(JB):                         0.00
Kurtosis:                    1120.050   Cond. No.                         192.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Choosing Sales as target and Shipping Cost and Profit as features, we can create the model to predict Sales. Above, Adjusted R-Squares and R-Squares are same. By checking Shipping Cost and Profit variables, 60% of Sales can be predicted. The model validates the rejection of null hypothesis.



	Coeffecient
Shipping Cost	5.909045
Profit	0.560526



Conclusion

This analysis has shown that Shipping Cost and Profit are statistically significant values to determine Sales. In order to increase Sales, the company needs to focus on Shipping Cost more than any other variables. Discount and Quantity are much more an influence rather than significant values on Sales. Company should sell more technology products rather than other categories due to high profitability. Especially furniture category has almost no significant profitability to company. Although company can still sell furniture and office supplies, price increasement in these two categories is essential.

References

Chen. E., 2019 Time Series Analysis on Super Store Sales Data Retrieved from https://haochen23.github.io/2019/02/time-series-analysis-superstore-sales.html#.XyyQGyhKjIU

Creating and Updating Figures in Python Retrieved from on August 8 from https://plotly.com/python/creating-and-updating-figures/

EDA-Super Store Data Retrieved on August 8 from https://www.kaggle.com/shreyashitiwari/eda-superstore-data

Seaborn boxplot Retrieved on August 8 from https://seaborn.pydata.org/generated/seaborn.boxplot.html

Sns Boxplot Retrieved on August 8 from https://www.kaggle.com/ashydv/sales-prediction-simple-linear-regression

Super Store Assignment Final Retrieved on August 8 from https://www.kaggle.com/pealdasgupta/superstore-assignment-final

Super Store Sales Data Retrieved on August 8 from https://www.kaggle.com/jr2ngb/superstore-data

Time Series Forecasting of Super Store Data Set Analysis Retrieved on August 8 from https://github.com/vibhor98/Time-Series-Forecasting-of-Superstore-dataset/blob/master/Analysis.ipynb