**Superstore**

**Exploratory Analysis**

Hannah Garrett, [hgarrett@bellarmine.edu](mailto:hgarrett@bellarmine.edu)

McKenzee Masters, [mmasters@bellarmine.edu](mailto:mmasters@bellarmine.edu)

1. **INTRODUCTION**

With growing demands and cu-throat competitions in the market, this superstore data set allows the user to analyze different aspects within a market in order to understand what will work best for this hypothetical Superstore. This data set helps the user understand which products, regions, categories, and customer segments they should target or avoid. You can also use this data to predict sales and profits for the superstore. As business majors/minors, this data set gives a glimpse of how data will be analyzed in the real business world.

1. **DATA SET DESCRIPTION**

Narrative summary of the data set: e.g. this data set contains 398 samples with 7 columns with various data types. A complete listing is shown in **Table 1**. For data types you want to indicate two things (nominal, ordinal, interval, or ratio) and the Pandas data type. For example, age might be ratio/int32. For missing data, indicate what percentage of data from that column are missing. Ensure you check to for NaN, NA, or any other indicators that actually mean missing data.

This data set contains 9994 samples with 21 columns of various data types. There was no missing data in this set.

A complete listing is shown in **Table 1**.

**Table 1: Data Types and Missing Data**

|  |  |  |
| --- | --- | --- |
| *Variable Name* | *Data Type* | *Missing Data (%)* |
| Row ID | Int64 | 0% |
| Order ID | Object | 0% |
| Order Date | object | 0% |
| Ship Date | object | 0% |
| Ship Mode | object | 0% |
| Customer ID | object | 0% |
| Customer Name | object | 0% |
| Segment | object | 0% |
| Country | object | 0% |
| City | object | 0% |
| State | object | 0% |
| Postal Code | Int64 | 0% |
| Region | object | 0% |
| Product ID | object | 0% |
| Category | object | 0% |
| Sub-Category | object | 0% |
| Product Name | object | 0% |
| Sales | Float64 | 0% |
| Quantity | object | 0% |
| Discount | Float64 | 0% |
| Profit | Float64 | 0% |

1. **Data Set Summary Statistics**

Here we will see the summary of statistics for the numerical variables within the Superstore data set. We will also look at the proportions for categorical variables, as well as looking at the correlations.

**Table 2: Summary Statistics for Superstore**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Variable Name* | *Count* | *Mean* | *Standard Deviation* | *Min* | *25th* | *50th* | *75th* | *Max* |
| Sales | 9994 | 229.85 | 623.24 | 0.444 | 17.28 | 54.49 | 209.94 | 22638.48 |
| Quantity | 9994 | 3.78 | 2.225110 | 1 | 2 | 3 | 5 | 14 |
| Discount | 9994 | 0.15 | 0.206452 | 0 | 0 | 0.2 | 0.2 | 0.8 |
| Profit | 9994 | 28.65 | 234.260108 | -6599.9 | 1.72875 | 8.6665 | 29.364 | 8399.976 |

There should be a table for **EACH** categorical variable.

Table 3: Proportions for Order Year

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| 2017 | 3312 | 33.14 |
| 2016 | 2587 | 25.89 |
| 2015 | 2102 | 21.03 |
| 2014 | 1993 | 19.94 |

Table 4: Proportions for Order Month

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| 1 | 381 | 3.81 |
| 2 | 300 | 3.00 |
| 3 | 696 | 6.96 |
| 4 | 668 | 6.68 |
| 5 | 735 | 7.35 |
| 6 | 717 | 7.17 |
| 7 | 710 | 7.10 |
| 8 | 706 | 7.06 |
| 9 | 1383 | 13.83 |
| 10 | 819 | 8.19 |
| 11 | 1471 | 14.72 |
| 12 | 1408 | 14.09 |

Table 5: Proportions for Customers

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| William Brown | 37 | 0.37 |
| John Lee | 34 | 0.34 |
| Matt Abelmen | 34 | 0.34 |
| Paul Prost | 34 | 0.34 |
| Chloris Kastensmidt | 32 | 0.32 |

Table 6: Proportions for City

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| New York City | 915 | 9.15 |
| Los Angeles | 747 | 7.47 |
| Philadelphia | 537 | 5.37 |
| San Francisco | 510 | 5.10 |
| Seattle | 428 | 4.28 |

Table 7: Proportions for States

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| California | 2001 | 20.02 |
| New York | 1128 | 11.28 |
| Texas | 985 | 9.85 |
| Pennsylvania | 587 | 5.87 |
| Washington | 506 | 5.06 |

Table 8: Proportions for Region

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| West | 3203 | 32.05 |
| East | 2848 | 28.49 |
| Central | 2323 | 23.24 |
| South | 1620 | 16.21 |

Table 9: Proportions for Product

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| Staple Envelope | 48 | 0.48 |
| Staple | 46 | 0.46 |
| Easy-staple paper | 46 | 0.46 |
| Avery non-stick binders | 20 | 0.20 |
| Staples in misc. colors | 506 | 5.06 |

Table 10: Proportions for Category of Product

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| Office Supplies | 6026 | 60.29 |
| Furniture | 2121 | 21.22 |
| Technology | 1847 | 18.48 |

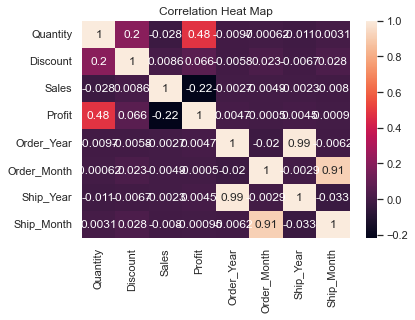
Table 11: Proportions for Region

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| Binders | 1523 | 15.24 |
| Paper | 1370 | 13.71 |
| Furnishing | 957 | 9.57 |
| Phones | 889 | 8.89 |
| Storage | 846 | 8.46 |

After you summarize the categorical variables, generate a correlation matrix for all continuous variables (not categorical – this doesn’t make sense)

Table 12: Correlation Table/Tables

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Quantity | Discount | Sales | Profit | Year Ordered | Month Ordered | Year Shipped | Month Shipped |
| Quantity | 1.000000 | 0.008623 | 0.200795 | 0.066253 | -0.005810 | 0.023355 | -0.006716 | 0.027664 |
| Discount | 0.008623 | 1.000000 | -0.028190 | -0.219487 | -0.002662 | -0.004935 | -0.002279 | -0.008044 |
| Sales | 0.200795 | -0.028190 | 1.000000 | 0.479064 | -0.009679 | -0.000625 | -0.010815 | 0.003110 |
| Profit | 0.066253 | -0.219487 | 0.479064 | 1.000000 | 0.004670 | -0.000502 | 0.004531 | -0.000951 |
| Year Ordered | -0.005810 | -0.002662 | -0.009679 | 0.004670 | 1.000000 | -0.020183 | 0.994083 | -0.006160 |
| Month Ordered | 0.023355 | -0.004935 | -0.000625 | -0.000502 | -0.020183 | 1.000000 | -0.002905 | 0.911882 |
| Year Shipped | -0.006716 | -0.002279 | -0.010815 | 0.004531 | 0.994083 | -0.002905 | 1.000000 | -0.033411 |
| Month Shipper | 0.027664 | -0.008044 | 0.003110 | -0.000951 | -0.006160 | 0.911882 | -0.033411 | 1.000000 |

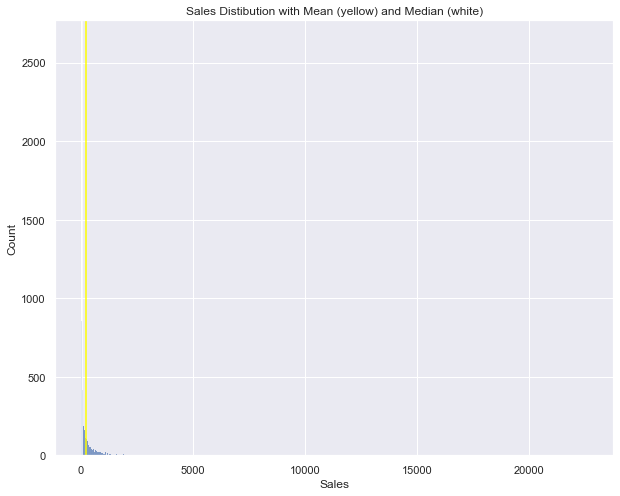


1. **DATA SET GRAPHICAL EXPLORATION**

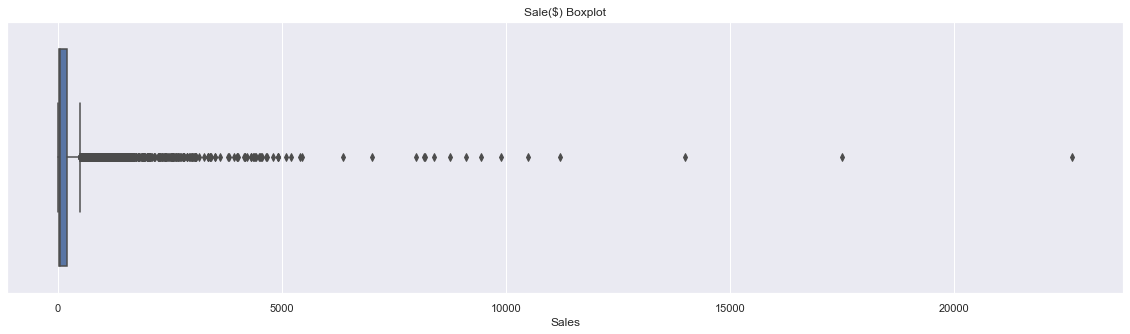
Narrative introduction to the section. In each section below, indicate any interesting distributions, anomalies, imbalance, etc. that you notice.

**Sales**

The distribution of sales is extremely right skewed with many outliers. Since there are so many outliers, the distribution is very hard to read on the histogram (Figure 1) since the number of sales ranges from 0 to 22000 dollars. . From the box plot (Figure 2), we can see the various outliers. The yellow line represents the mean of $228.86. Out of a 9,994 sample size, most values fell below $2000.

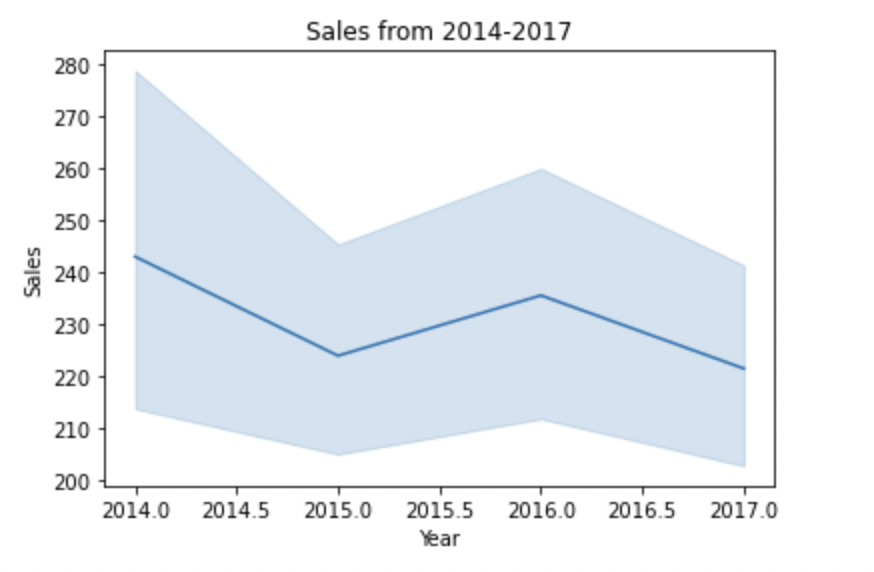


**Figure 1: Distribution of number of Sales in dollars 2014-2017**



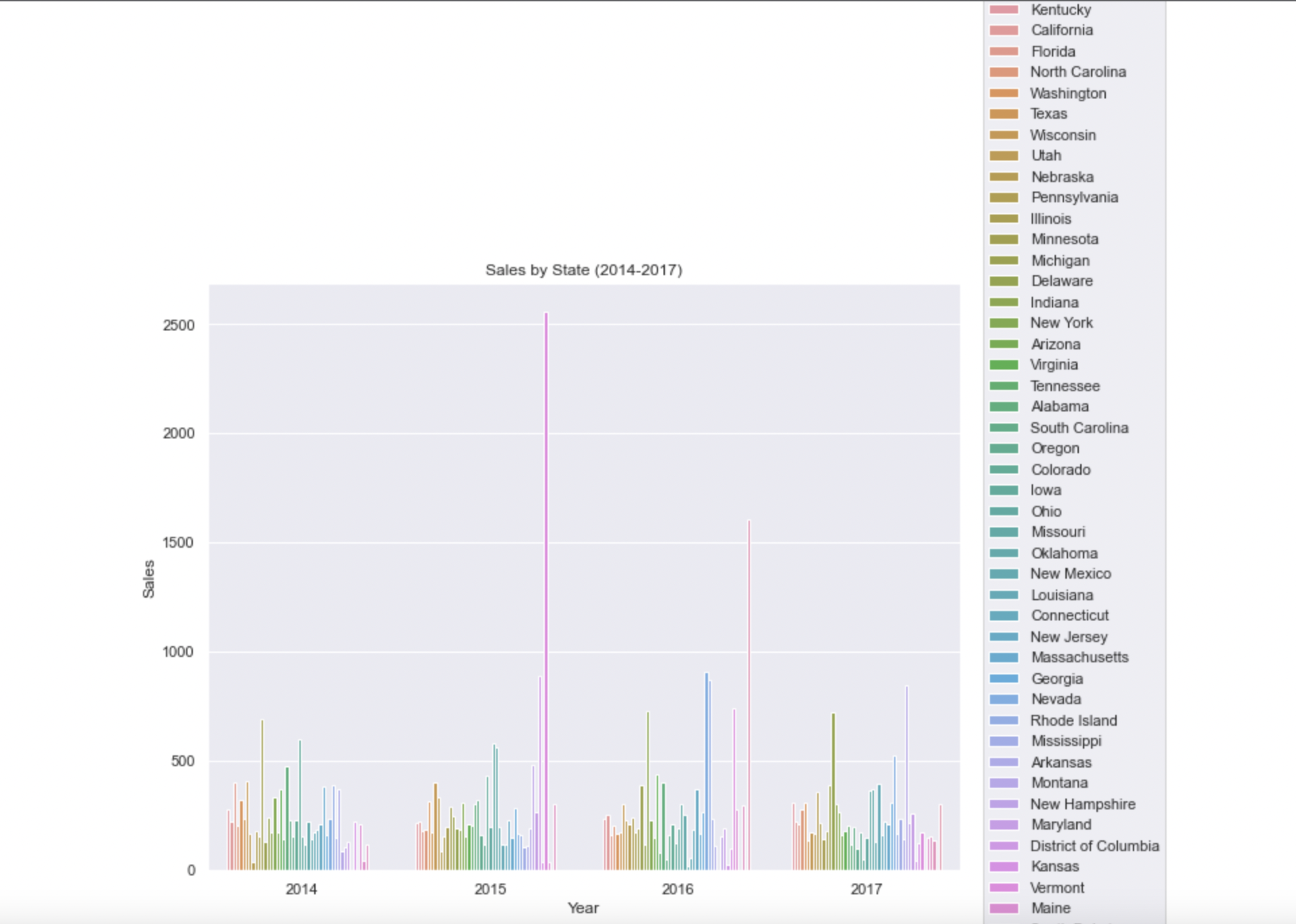
**Figure 2: Boxplot of Sales including outliers**

From the sales data, we also analyzed the trend of sales over the years. From the line plot in figure 3, we can see the number of sales decreased from 2014 to 2017.



**Figure 3: Trend of sales from 2014 to 2017**

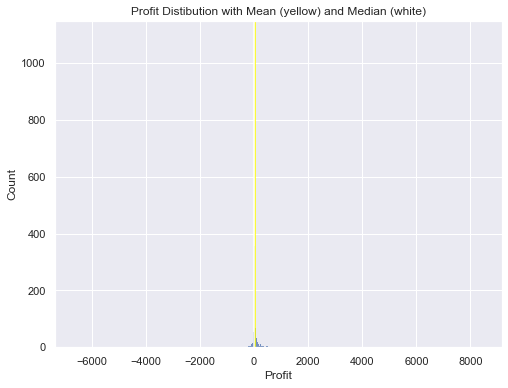
The bar plot in figure 4 represents the number of sales by state. We can see that states generally had the most sales in 2015 and 2016.



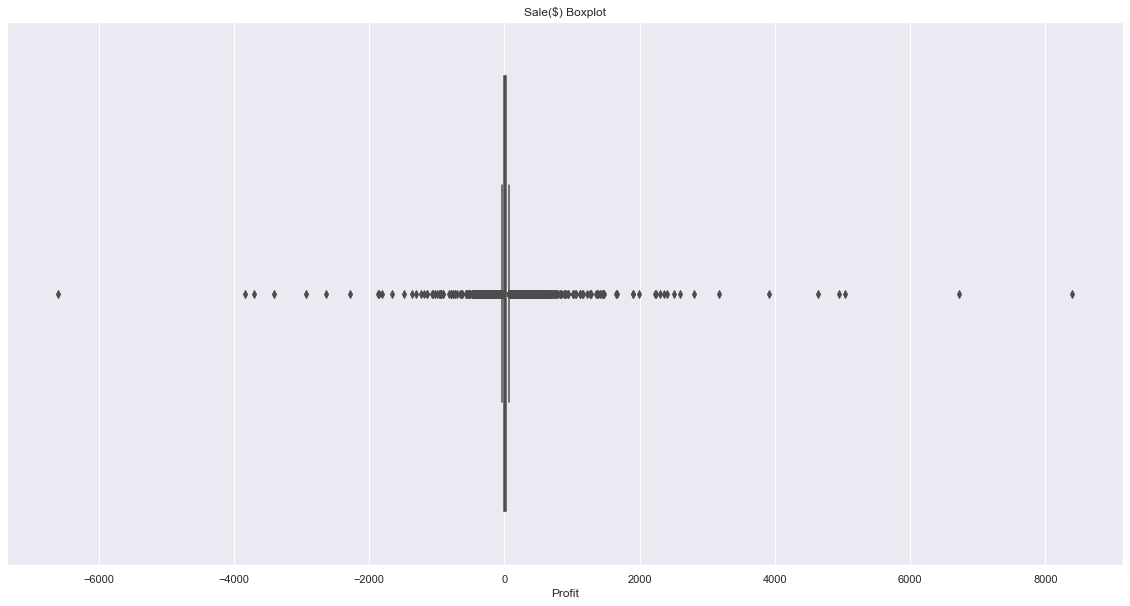
**Figure 4: Number of Sales by State**

**Profits**

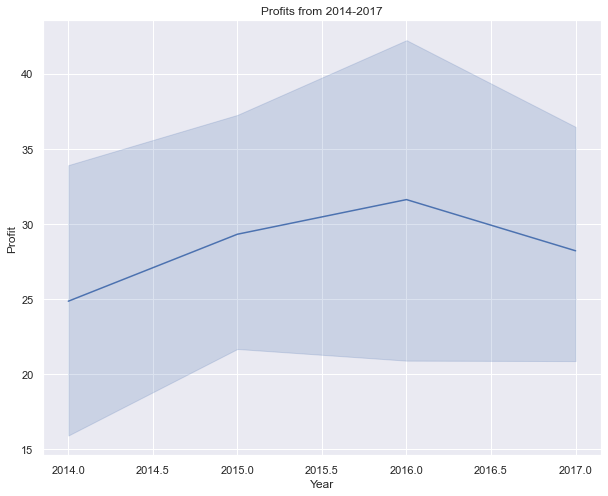
The distribution of profits is very similar to the distribution of sales. Figures 5 and 6 are very similar to figures 1 and 2. Figure 5 shows that the distribution of profits is normal, but there are so many outliers that it makes the graph difficult to read. We can see the outliers in figure 6. The trend of profits increases until 2016, where profits start to decrease (Figure 7).



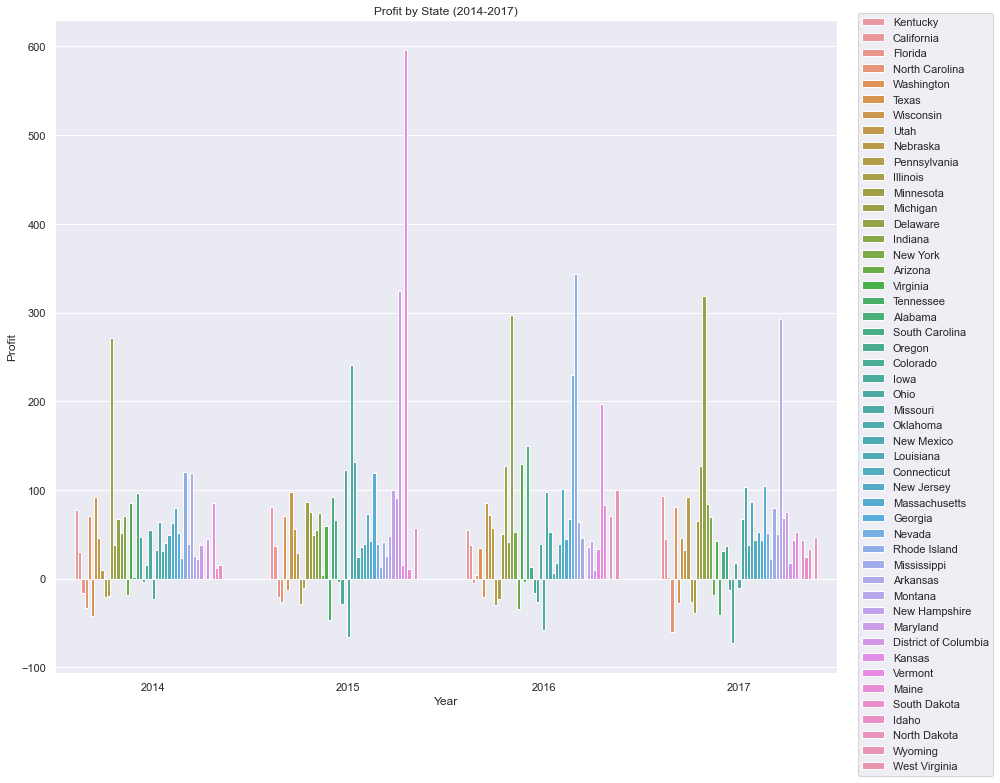
**Figure 5: Histogram distribution of count of Profit**



**Figure 6: Box plot of profits**



**Figure 7: Line plot depicting trend of profits 2014-2017**



**Figure 8: Profits by State**



Figure 9: Quantity of Item Bought vs Profit



Figure 10: Quantity of Item Bought vs Sales

1. **SUMMARY OF FINDINGS**

While investigating the Superstore data set, we found correlations between quantity of items bought and profit, order year and ship year, and order month and ship month. The correlation between quantity and profit shows that the Superstore makes more profit on items that can be bought in higher quantities, like paper and staples. The correlation between order and ship year and order and ship month shows that the Superstore has a fast turnaround for order processing.