Final Assessment - COMP8060

Quentin JONNEAUX - R00274704

2024-12-21

Introduction

In this report, we are helping a global distribution organization to gather to provide indepth analysis of their data. Especially, the company is interested in understanding aspects such as sales, shipping costs and what are the factors affecting them according to the data (and indeed factors that may have been assumed to affect them but don't according to the data).

We are tasked with the production of an application that allows a user in the organisation to explore some of the most interesting aspects of this dataset. We will compute the dataset and implement Unit Profits statistics.

We will then assess the evolution of sales, profits and quantities ordered over the years, by providing line plots and barplots, before interpreting those.

After that, we will try to understand what is affecting the Shipping Costs. We will produce a correlation matrix among numeric data and provide a plot checking the impact of Shipping Mode on cost.

Thus, we will compute the option for the user to input 2 countries for comparison. This comparison will provide the value counts for subcategories and visualisations for the comparison.

We will also give our personal insights on the dataset. It is always appreciated to have external point of view on the data to challenge assumptions. We will provide visualisations and comments regarding those.

Finally, we will give a summary and recommendations on our findings.

Insights provided by the company

We are provided a csv file containing a dataset. The dataset is comprised of 50631 rows and the following 22 columns:

- category: The category of the product.
- sub category: The sub-category of products within the main category.
- segment: The customer segment (e.g., consumer, corporate, or home office).
- city: The city where the order was placed.
- state: The state or region within the country.
- country: The country in which the store is located.
- region: The region where the order was placed.
- market: The market or region where the store operates. • order date: The date when the order was placed.
- order id: A unique identifier for each order
- order priority: The priority level of the order.
- product id: A unique identifier for each product. • product name: The name of the product.
- profit: The profit generated from the order.
- quantity: The quantity of products ordered.
- discount: The discount applied to the order.
- sales: The total sales amount for the order.
- ship mode: The shipping mode used for the order.
- shipping cost: The cost of shipping for the order.
- ship date: The date when the order was shipped.
- shipM: The month when the order was shipped.
- shipY: The year when the order was shipped.

We can notice a significant amount of missing values spread in different columns, so data cleaning will be necessary while computing functions.

1 – Main function, computing Unit Profits and summary of the dataset

The program will prompt the user to use a single number between 1 and 6 to be provided with statistics and visualisations. Here are the choices:

- 1 – Initial Data Summary
- 2 – Sales Analysis
- 3 – Shipping Cost Analysis
- 4 – Country comparison for Top Product Categories
- 5 – Personal Insights
- 6 – Exit the program

The initial data summary also computes a new variable in the data set. This Unit Profit variable is added to the data and is the profit divided by the quantity. The summary is broken down in 3 parts:

- Descriptive statistics of numerical variables
- Descriptive statistics of non-numerical variables
- The first 5 rows of the computed data set

Descriptive statistics of numerical variables

9 variables in the data set are numeric. Let's see the statistics outputted:

Please find the numeric descriptive statistics of the dataframe:

	Profit	Quantity	Discount	Sales	Shipping.Cost
count	50306.000000	50484.000000	50176.000000	50457.000000	50165.000000
mean	28.639829	3.475517	0.142933	246.755435	26.389863
std	174.637918	2.278279	0.212363	488.822000	57.308674
min	-6599.978000	1.000000	0.000000	0.000000	0.002000
25%	0.000000	2.000000	0.000000	31.000000	2.610000
50%	9.251000	3.000000	0.000000	85.000000	7.790000
75%	36.810000	5.000000	0.200000	251.000000	24.460000
max	8399.976000	14.000000	0.850000	22638.000000	933.570000

	weeknum	ShipY	ShipM	UnitProfit
count	50631.000000	50631.000000	50631.000000	50160.000000
mean	31.163220	2012.780293	7.499003	8.104029
std	14.377605	1.095932	3.319892	43.383342
min	1.000000	2011.000000	1.000000	-1319.995600
25%	20.000000	2012.000000	5.000000	0.000000
50%	33.000000	2013.000000	8.000000	3.420000
75%	44.000000	2014.000000	10.000000	12.251100
max	53.000000	2014.000000	12.000000	1679.995200

Since the dataset is comprised of 50631 rows, we know that all variables, except "weeknum", "ShipY" and "ShipM" are containing missing values.

The "Profit" mean is 28.639829 but also has a significant standard deviation. We also notice an IQR of 36.81 and a minimum and maximum value much outside 1.5 IQR on each side, so we have to keep in mind there are outliers present. Median is 9.251.

The "Quantity" mean is 3.475517 but has a less significant standard deviation. We also notice an IQR of 3 and a maximum value of 14 (much outside 1.5 IQR above), so we have to keep in mind there are outliers present. Median is 3.

The “Discount” mean is 0.142933 and standard deviation is 0.212363. We also notice an IQR of 0.2 and a maximum value of 0.85 (much outside 1.5 IQR above), so we have to keep in mind there are outliers present. Median is 0.

The “Sales” mean is 246.755435 and standard deviation is 488.822. We also notice an IQR of 220, a minimum value of 0 and a maximum value of 22638 (much outside 1.5 IQR above), so we have to keep in mind there are outliers present. Median is 85.

The “Shipping Cost” mean is 26.389863 and standard deviation is 57.308674. We also notice an IQR of 21, a minimum value of 0.002 and a maximum value of 933.57, so we have to keep in mind there are outliers present. Median is 7.79.

The “Unit Profit” mean is 8.104029 and standard deviation is 43.383342. We also notice an IQR of 122511 and a maximum value of 1679.9952 (much outside 1.5 IQR above), so we have to keep in mind there are outliers present. Median is 3.42.

The other numeric variable implies time and do not contain missing values. We can see sales are recorded between 2011 and 2014 (inclusive).

Descriptive statistics of non-numerical variables

15 variables in the data set are non-numeric. Let’s see the statistics outputted:

```
Please find the categorical descriptive statistics of the dataframe:

              Category Sub.Category   Segment           City       State
count            50514        50360     50207          50387       50218
unique               3           17         3           3622        1090
top     Office Supplies      Binders  Consumer  New York City  California
freq             30796         6040     25976            898        1964

             Country   Region Market              Order.Date  \
count          50226    50222  50410                   50430
unique           147       13      7                    1429
top     United States  Central   APAC  2014-06-18 00:00:00.000
freq            9796    10876  10828                     135

             Order.ID Order.Priority       Product.ID Product.Name  \
count           50331          50499            50344        50601
unique          24725              4            10265         3787
top     CA-2014-100111         Medium  OFF-AR-10003651      Staples
freq               14          28965               35          224

            Ship.Mode               Ship.Date
count           50170                   50631
unique              4                    1457
top     Standard Class  2014-11-22 00:00:00.000
freq            30057                     130
```

Since the dataset is comprised of 50631 rows, we know that all variables, except “Ship.Date” are containing missing values.

The “Category” comprises 3 unique categories where Office supplies records the highest frequency.

The “Sub.Category” comprises 17 unique categories where Binders records the highest frequency.

The “Segment” comprises 3 unique categories where Consumers records the highest frequency.

“City” and “State” contains many different categories. The “Country” comprises 147 unique categories where United States records the highest frequency.

The “Region” comprises 13 unique categories where Central records the highest frequency.

The “Market” comprises 7 unique categories where APAC records the highest frequency. Each location highest frequencies do not belong to the same continent (Country, Region, Market), it will be interesting to see if location affects the variables.

The “Ship.Mode” comprises 4 unique categories where Standard class records the highest frequency.

The other categories are very diverse and allow to identify sales to time, products and priorites.

The first 5 rows of the computed data set

Please find the first 5 rows of the dataframe:

	Category	Sub.Category	Segment	City	State
0	Office Supplies	Paper	Consumer	Los Angeles	California
1	Office Supplies	Paper	Consumer	Los Angeles	California
2	Office Supplies	Paper	Consumer	Los Angeles	California
3	Office Supplies	Paper	Consumer	Los Angeles	California
4	Office Supplies	Paper	Consumer	Los Angeles	California

	Country	Region	Market	Order.Date	Order.ID
0	United States	West	US	2011-01-07 00:00:00.000	CA-2011-130813
1	United States	West	US	2011-01-21 00:00:00.000	CA-2011-148614
2	United States	West	US	2011-08-05 00:00:00.000	CA-2011-118962
3	United States	West	US	2011-08-05 00:00:00.000	CA-2011-118962
4	United States	West	US	2011-09-29 00:00:00.000	CA-2011-146969

	Order.Priority	Product.ID
0	High	OFF-PA-10002005
1	Medium	OFF-PA-10002893
2	Medium	OFF-PA-10000659
3	Medium	OFF-PA-10001144
4	High	OFF-PA-10002105

	Product.Name	Profit	Quantity
0	Xerox 225	9.3312	3.0
1	Wirebound Service Call Books, 5 1/2" x 4"	9.2928	2.0
2	Adams Phone Message Book, Professional, 400 Me...	9.8418	3.0
3	Xerox 1913	53.2608	2.0
4	Xerox 223	3.1104	1.0

	Discount	Sales	Ship.Mode	Shipping.Cost	Ship.Date
0	0.0	19.0	Second Class	4.37	2011-01-09 00:00:00.000
1	0.0	19.0	Standard Class	0.94	2011-01-26 00:00:00.000
2	0.0	21.0	Standard Class	1.81	2011-08-09 00:00:00.000
3	0.0	111.0	Standard Class	4.59	2011-08-09 00:00:00.000
4	0.0	6.0	Standard Class	NaN	2011-10-03 00:00:00.000

	weeknum	ShipY	ShipM	UnitProfit
0	2	2011.0	1.0	3.1104
1	4	2011.0	1.0	4.6464
2	32	2011.0	8.0	3.2806
3	32	2011.0	8.0	26.6304
4	40	2011.0	10.0	3.1104

Main insights

This summary is quite extensive so here are the main insights:

- There are outliers in every variable of interest (Profits, Quantity, Discount, Sales, Shipping Costs and Unit Profits)
- Negative profits has been recorded (minimum value of -6599.978) are centralized between 0 and 36.81 (Median 9.251). It distribution seems right skewed.

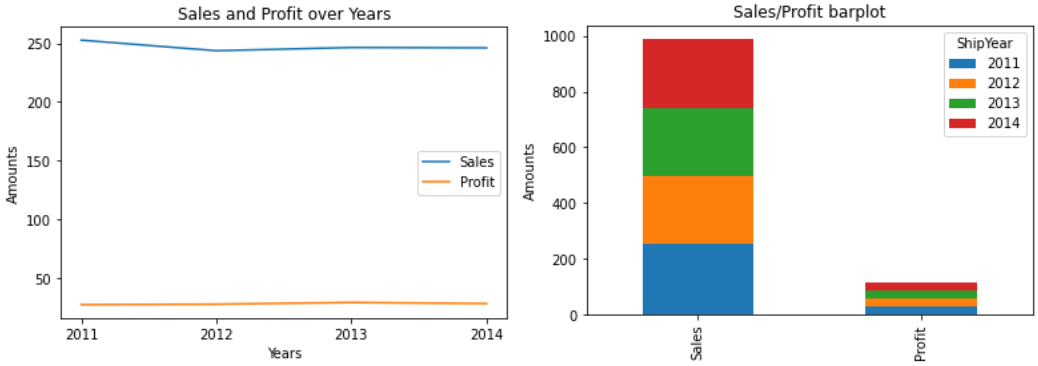
- Quantities seems to follow an approximate normal distribution (Mean is 3.475517, Median 3) and usually varies between 1 and 5.
- Half of the data does not record any discounts
- Sales seems to follow a right skewed distribution (Mean is 246.755435, Median 85).
- Shipping Costs seems to follow a right skewed distribution (Mean is 26.389863, Median 7.79).
- Unit Profit seems to follow a right skewed distribution (Mean is 8.104029, Median 3.42)
- The set contains 17 Subcategories of Products in 3 segments of customer in 147 different countries, using 4 different Shipping Modes.

2 – Sales Analysis

The variable of interest for the sales analysis are the profits, the sales and the quantities.

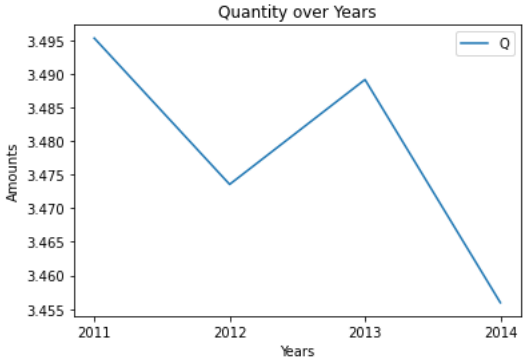
Yearly evolution

Let's see how the sales and profits are evolving over the years and how they a distributed:



We can see from the line plot that that sales slightly dropped from 2012 but are evolving steadily over the years around 250. Profits are also steady over the years around 20. The stacked bar plot suggest that both Sales and Profit looks evenly distributed across years (25% for each year).

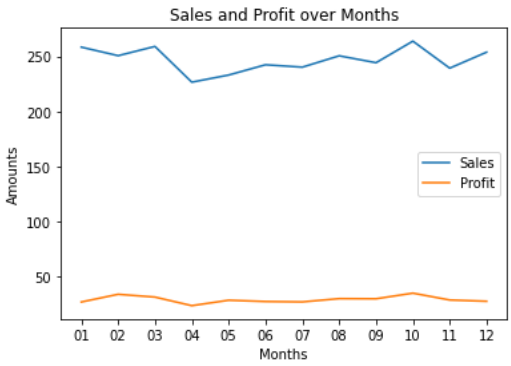
Let's see the quantities over the years:

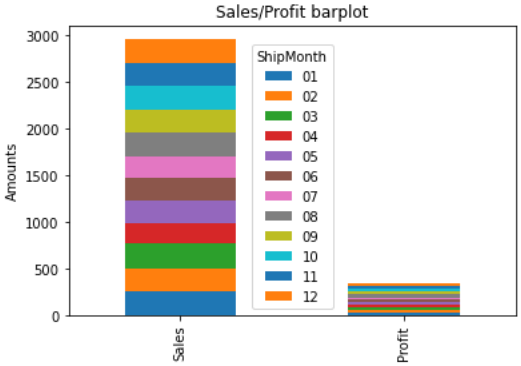


The quantities fluctuate from 3,495 to 3,455, with a decrease to 3,475 in 2012, then an increase to 3,490 in 2013. When put side by side, it is interesting to see while sales and profits look steady, quantities are decreasing over years. Over years, quantities do not seem to affect sales or profit.

Monthly evolution

Let's see how the sales and profits are evolving over the years and how they a distributed:





We can see from the lineplot that that sales fluctuate over months with a minimum value in April and the maximum value in October. Profits are also steady but peaks are not as noticeable in sales. The stacked barplot suggest that both Sales and Profit looks evenly distributed across months.

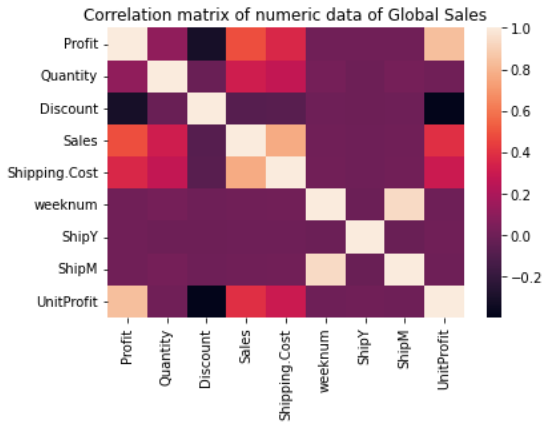
Let's see the quantities over the months:

The quantities increase over months, hitting the lowest value of 3,375 in February and highest value of 3,575 in October. September-October has the biggest difference in Quantities. When put side by side, it is interesting to see while sales and profits look steady, quantities are increasing over months. Over years, quantities do not seem to affect sales or profit, however highest values of Quantity and Profits seems to match on the same month (October).

3 – Shipping Costs Analysis

Let's figure out what is affecting the shipping cost. It is assumed that shipping mode affects the shipping costs. Let's generate a heatmap of correlations to analyse the numeric data and boxplots to analyse the distributions of cost according to mode.

Heatmap of Correlations



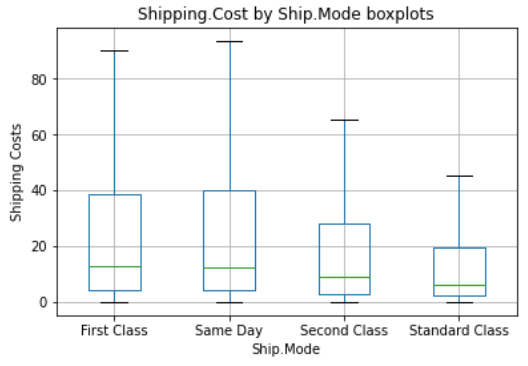
This heatmap quantifies the impact of numeric variables. We can see that Shipping Costs are not affected by the date (weeknum, ShipY and ShipM).

We can see that there is a correlation between Shipping Costs and Sales (0.8).

Then comes Profit and Unit Profit as affecting factors of Shipping Cost (0.5), and Quantity (0.2).

Finally, the discounts have a very minor impact on Shipping Costs (-0.1) but it is the only one affecting it negatively.

Boxplots of Shipping Mode



The boxplots show the distributions of Shipping Costs with the Shipping Modes.

We removed outliers as numerous are present in each category and it would be difficult to provide readable visualisation. We can see all are right skewed but the lowest median and narrowest class is the standard class. It is the less spread class in Range and IQR.

The second class has a slightly higher median and a slightly larger IQR and range.

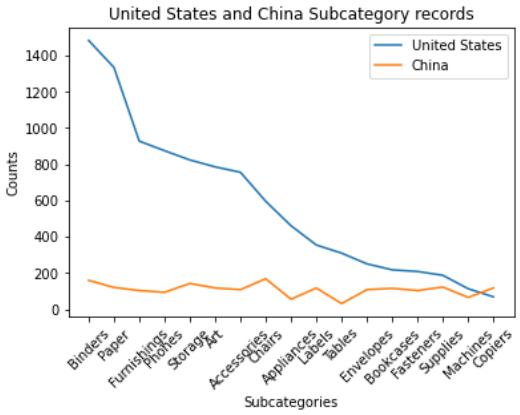
The First class and same day class have slightly higher medians but significantly larger range and IQR.

In other words, Shipping Mode has a little effect on medians but a significant effect on spread. We can give the following ranking from less impact to much impact:

- Standard Class
- Second Class
- First Class
- Same Day

4 – Country Comparison for Top Products

It is of interest to see the value counts of subcategories between 2 countries. We noticed in Descriptive Statistics that Country and Market max counts were not aligned (United States and APAC). Let's make a comparison between US and an APAC Country, China. Let's plot the value counts to compare distributions in common subcategories:

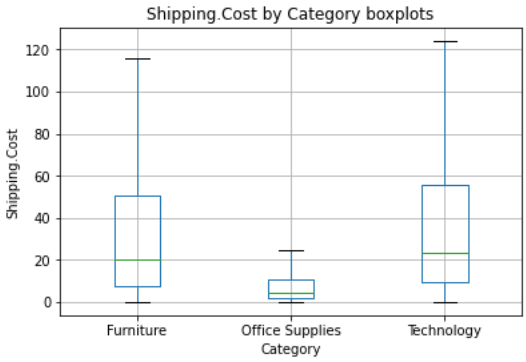


We can see distributions are very different. While China seems to have a uniform distribution across subcategories (around 100 and 200 in each subcategory), US is not uniform. While some categories are alike China's (Supplies, Machines, Copiers), some categories are recording much higher value counts (Binders, Paper, Furnishing).

Based on this plot, we can say that country is affecting the distribution of sales of the subcategories.

5 – Personal Insights

We noticed that Shipping Cost are affected by shipping mode. It would be interesting to see if they are affected by categories.



We also removed outliers to make the visualization readable. It seems Furniture and Technology categories are similar in range, median and IQR. Distribution of Shipping Cost are similar (right skewed). But the distribution of Office supplies is much different. Although also right skewed, the range and IQR a much narrower and the median is significantly lower. We can assume that Office supplies categories reduce shipping cost significantly.

It would be also interesting to see the relationships between Shipping cost and other numerical variables. We create a heatmap in a previous section and let's plot each value to see if we can visualize the central location and spread of data.

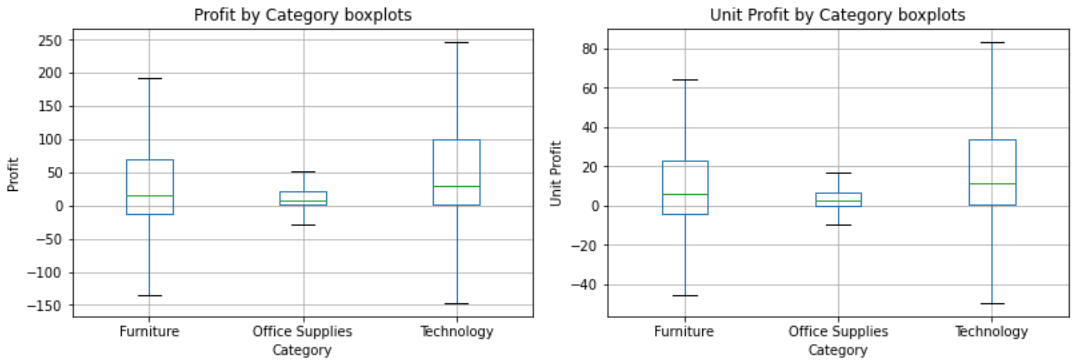
We can see that Sales have a slightly positive linear relationship with sales and most of the data is located under 4,000 sales and 500 of shipping costs. We can also notice that the range of sales narrows with shipping costs. Less sales are done if the shipping cost are too high.

We can see Profits and Unit Profit have a similar constant relationship with Shipping Costs. Data seems centralized under 500 of Shipping Costs. Profits seems centralized between -2000 and 2000 and Unit Profit between -500 and 500. The spread of data looks similar. We can argue that they do not have a strong relationship.

The quantity aspect seems to be another story.

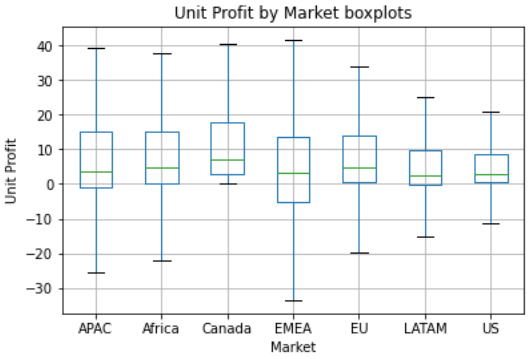
We can see that Shipping Cost have a strong positive relationship with Quantities, meaning the more quantities ordered, the more Shipping Costs. We can see a steady relationship until 9 order and then data fluctuates. When ordering large quantities, it would be sometimes advantageous to order more than 10, but we can be sure that the shipping cost steadily increase until 9.

Since Shipping Cost are affected by categories, it would be interesting if Profit has a similar pattern.



Indeed, we can see that the distributions of Furniture and Technology are similar while the Office supplies is narrower in range and IQR. However, while both Office supplies and Technology IQR are positive, it seems a significant proportion of the lower quartile of Furniture is in the negative profit, meaning a significant part of Furniture sales are not profitable. Medians for all categories are still positive and Technology seems to be the most positively distributed (highest median of profit and widest IQR). We can see the Unit Profit follow the same pattern, meaning adjust the quantity per sale will not affect the profits distribution.

Finally, since we would to understand profits, let's see if the Unit Profits are affected the target Market.



It seems Market is affecting Unit Profits as each market has a different distribution. Canada seems to be the most reliable since the range of value is positive, the higher quantile and the median are the highest. On the other hand, the range of the EMEA market is the widest and a significant part of the lower share of IQR is negative, meaning there is a significant share of unprofitable sales there. All the other markets are mostly positive and US seems to have the narrower range and IQR.

Summary

According to this program:

- There are outliers in every variable of interest (Profits, Quantity, Discount, Sales, Shipping Costs and Unit Profits)
- Negative profits has been recorded (minimum value of -6599.978) are centralized between 0 and 36.81 (Median 9.251). It distribution seems right skewed.
- Quantities seems to follow an approximate normal distribution (Mean is 3.475517, Median 3) and usually varies between 1 and 5.
- Half of the data does not record any discounts
- Sales seems to follow a right skewed distribution (Mean is 246.755435, Median 85).
- Shipping Costs seems to follow a right skewed distribution (Mean is 26.389863, Median 7.79).
- Unit Profit seems to follow a right skewed distribution (Mean is 8.104029, Median 3.42)
- The set contains 17 Subcategories of Products in 3 segments of customer in 147 different countries, using 4 different Shipping Modes.
- Sales and Profits are steady over years and months, while Quantities are decreasing over years but increasing over months.
- Sales, Categories, Quantities and Shipping Mode affects Shipping Costs
- US distribution is not uniform over categories
- Unit profits have an important share in negatives in Furniture Category
- Unit profits have an important share in negatives in EMEA Market
- Canada seems the most reliable market in terms of Unit Profit

Recommendations:

- Investigate what is causing outliers in Profits, Quantity, Discount, Sales, Shipping Costs and Unit Profits
- Investigate if negative profits for Furniture have common grounds
- Work around Standard Class to reduce shipping cost.
- Investigate if negative profits on EMEA Market have common grounds.