Supply Chain Data Analytics

Stan Brouwer[1], Liz Chan[2], Maaike Lamberst[3], Niek Schroor[4]

[1]Vrije Universiteit,
[2]Master TSCM,
[3]Supply Chain Data analysis,
[4]Group 10,

Introduction

We analyze, forecast and interpret the Superstore sales provided by Tableau using different statistical and machine learning methods.

We describe our work in the PDF version. However, we would like to recommend reading our quarto manuscript *here* as it contains the **relevant** R code in the Article Notebook.

0.1 Data Pre-processing

The superstore data set we selected is of high quality. Thus we do the required data pre-processing, but included the hypothetical steps we would take were our data of lower quality to communicate our understanding of the data pre-processing process.

We took the following pre-processing steps:

- Improved column names by removing whitespaces
- Removed the Row_ID column as it can be inferred by it's index
- Removed all columns with a single unique value, as storing these would be redundant
- Ensured machine-readable date formats in yyyy-mm-dd as these usually differ per locale.
- Ensured proper decimal separators
- Calculated the number of missing values (both NA and empty string " ") per column.

```
[1] "None of the columns contains missing values"
```

Source: Article Notebook

After these steps (and transposing the table for better document formatting), the data looks as follows:

Table 1: First 5 Rows of the Data (Transposed)

Order_ID	CA-2016-152156	CA-2016-152156	CA-2016-138688
Order_Date	2016-11-08	2016-11-08	2016-06-12
Ship_Date	2016-11-11	2016-11-11	2016-06-16
Ship_Mode	Second Class	Second Class	Second Class
Customer_ID	CG-12520	CG-12520	DV-13045
Customer_Name	Claire Gute	Claire Gute	Darrin Van Huff
Segment	Consumer	Consumer	Corporate
City	Henderson	Henderson	Los Angeles
State	Kentucky	Kentucky	California
Postal_Code	42420	42420	90036
Region	South	South	West

Corresponding author: Stan Brouwer,

Product_ID	FUR-BO-10001798	FUR-CH-10000454	OFF-LA-10000240
Category	Furniture	Furniture	Office Supplies
Sub_Category	Bookcases	Chairs	Labels
Product_Name	Bush Somerset Collection Bookcase	Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back	Self-Adhesive Address Labels for Typewriters by Universal
Sales	261.96	731.94	14.62
Quantity	2	3	2
Discount	0	0	0
Profit	41.9136	219.5820	6.8714

Source: Article Notebook

There is some more processing to do, for instance the removal of outliers. However, by doing so we impose our own assumptions on the data. Let's start by evaluating the descriptive statistics of our data and check if further processing is required.

Table 2: Descriptive Statistics for Numeric Columns

Column	Min	Max	Mean	Median	StdDev
Postal_Code	1040	99301	55190.38	56430.5	32063.69
Sales	0.444	22638.48	229.858	54.49	623.2451
Quantity	1	14	3.789574	3	2.22511
Discount	0	0.8	0.1562027	0.2	0.206452
Profit	-6599.978	8399.976	28.6569	8.6665	234.2601

Table 3: Descriptive Statistics for Date Columns

Column	Earliest	Latest
Order_Date	2014-01-03	2017-12-30
Ship_Date	2014-01-07	2018-01-05

Source: Article Notebook

We inspected the orders with the lowest and highers price (Sales in USD). The most expensive orders were professional printers, camera's and teleconferencing units with high unit prices, and these orders often were of high Quantity. The orders with the lowest price where often binders, had a high Discount rate, and often a Quantity of just one.

We were fascinated by the orders with a negative profit. These all had high Discount rates, and often concerned the same items, such as the Cubify CubeX 3D Printer Triple Head Print. The orders with a negative Profit where often part of a larger order (for instance CA-2016-108196), and placed by customers that placed multiple orders. We suspect these negative Profit's to be caused by faulty items that receive discounts, general discount codes, or volume discounts. However, due to especially the high discounts on orders with negative profits, we assume these to be valid orders. This decision has also been influenced by the high quality of the data. As we found no missing values whats however, we suspect the chance of some weird but

valid orders to be higher than encountering mistakes here. *[this paragraph could use some rewriting]*

In figure x we plotted the sales of the most popular products. Unfortunately, the sales of individual products were too low to determine any meaningfull trends.

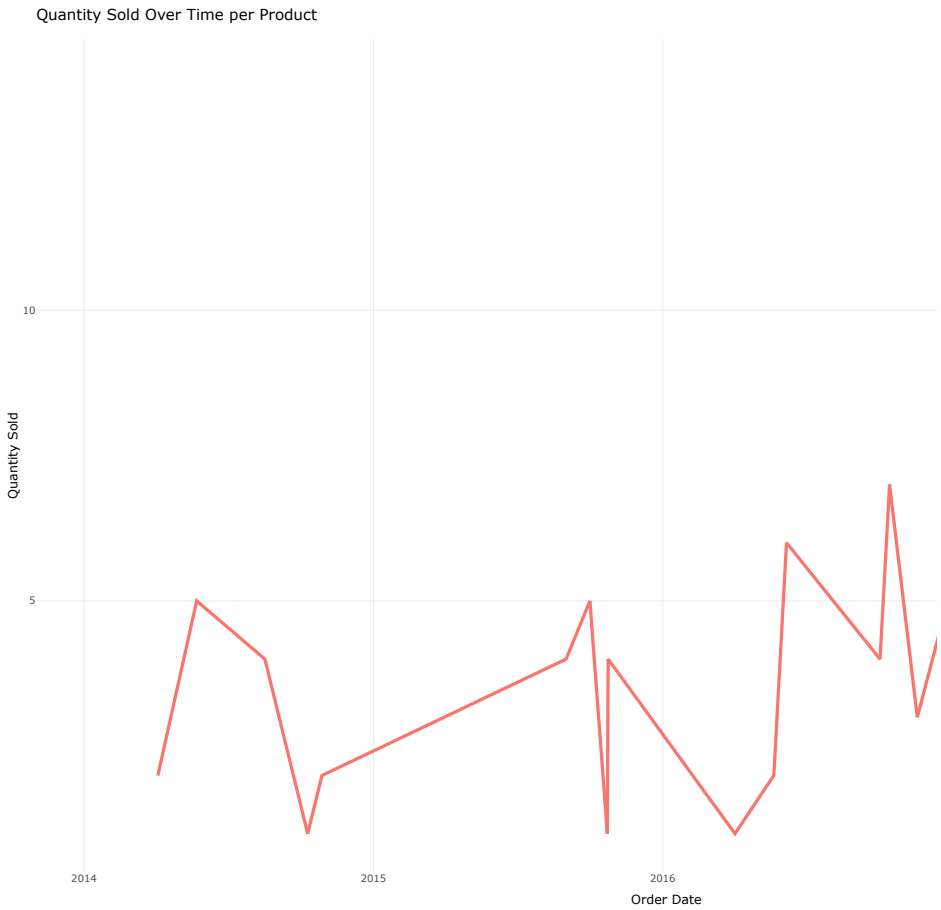
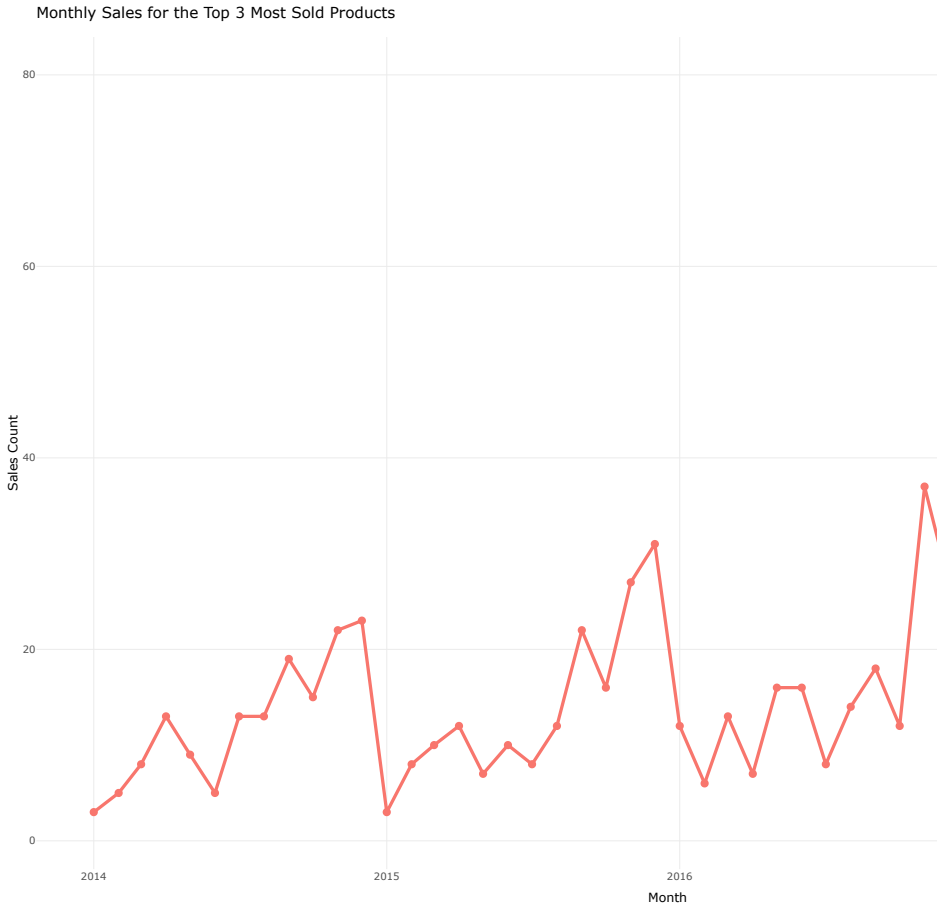


Figure 1: Figure x Sale quantity of the most popular products

Source: Article Notebook

Our proposed workaround is to aggregate products by their Sub_Category, and treating them as a single product for the rest of the assignment, which we plotted in figure X.



Source: Article Notebook

These aggregated sales start to show trends and seasonality, and are much more useful to base predictions on! We will use these aggregated sub-categories for the rest of the assignment.

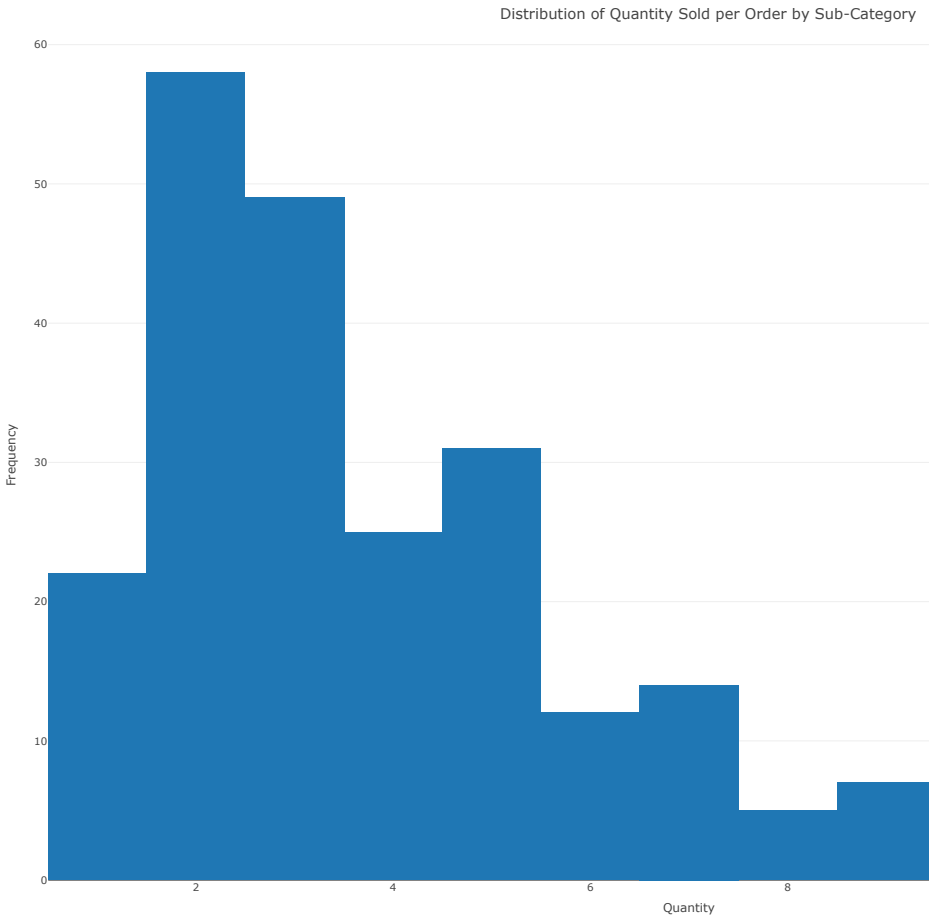
To properly finish our data pre-processing we ran some statistics on the aggregated sub-category sales. Table x contains soem descriptive statistics.

Table 4: Statistics for Sub_Category quantity

Sub_Category	Min	Mean	Max	Sd	CI_lower	CI_upper
Accessories	1	3.84	14	2.28	3.68	4.00
Appliances	1	3.71	14	2.12	3.52	3.90
Art	1	3.77	14	2.13	3.62	3.92
Binders	1	3.92	14	2.29	3.80	4.04
Bookcases	1	3.81	13	2.28	3.51	4.11
Chairs	1	3.82	14	2.28	3.64	4.00
Copiers	1	3.44	9	1.83	3.01	3.87
Envelopes	1	3.57	9	2.05	3.32	3.82
Fasteners	1	4.21	14	2.41	3.89	4.53
Furnishings	1	3.72	14	2.16	3.58	3.86
Labels	1	3.85	14	2.35	3.61	4.09
Machines	1	3.83	11	2.17	3.43	4.23
Paper	1	3.78	14	2.23	3.66	3.90
Phones	1	3.70	14	2.19	3.56	3.84
Storage	1	3.73	14	2.19	3.58	3.88
Supplies	1	3.41	10	1.84	3.15	3.67
Tables	1	3.89	13	2.45	3.62	4.16

65 Source: Article Notebook

66 The statistics for the sales aggregated by product category look valid. We can fur-
67 ther inspect them by visualizing them as histogram and visually check for anomalies.
68 Figure y contains histograms of the quantities per sub-category.



Source: Article Notebook

The histograms show that the quantities are not normally distributed, but have a right-skewed distribution. This is expected as most orders contain a small number of items, but some orders contain a large number of items. We will not remove these outliers as they are valid orders.

As the data we are going to use seems valid, we move on to exploring the trends and visualizing our data.

0.2 Data Visualization

some text for the visualization