Supply Chain Data Analytics

Analyzing and Forcasting Supermarket Sales

2024-12-05

Introduction

We analyze, forecast and interpret the [Superstore sales](https://public.tableau.com/app/sample-data/sample_-_superstore.xls) provided by [Tableau](https://public.tableau.com/app/learn/sample-data) 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.

## 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](https://few.vu.nl/~molenaar/courses/StatR/chapters/B-06-raw_data.html)
* 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](https://SJbrou.github.io/Supply_Chain_Data_Analysis/index.qmd.html)

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

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 |
| 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](https://SJbrou.github.io/Supply_Chain_Data_Analysis/index.qmd.html)

There is some more processing to do, for instance the removing 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.

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 |

Descriptive Statistics for Date Columns

| Column | Earliest | Latest |
| --- | --- | --- |
| Order\_Date | 2014-01-03 | 2017-12-30 |
| Ship\_Date | 2014-01-07 | 2018-01-05 |

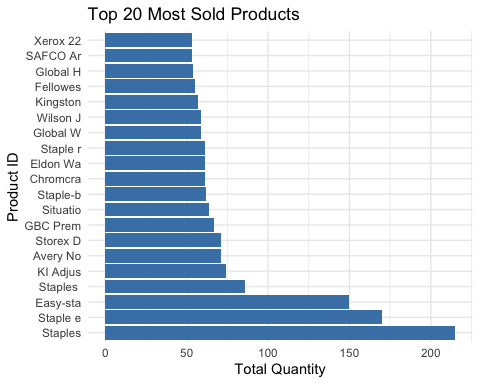
Source: [Article Notebook](https://SJbrou.github.io/Supply_Chain_Data_Analysis/index.qmd.html)

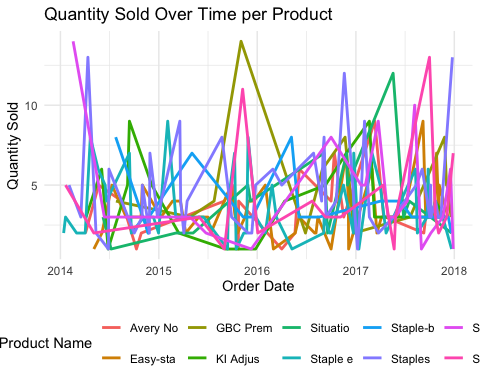
We inspected the orders with the lowest and highers price (Sales in USD). The most expensive orders were professional printers, camera’s and teleconferencing units with high unit prices, and these orders often were of high Quantity. The orders with the lowest price where often binders, had a high Discount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

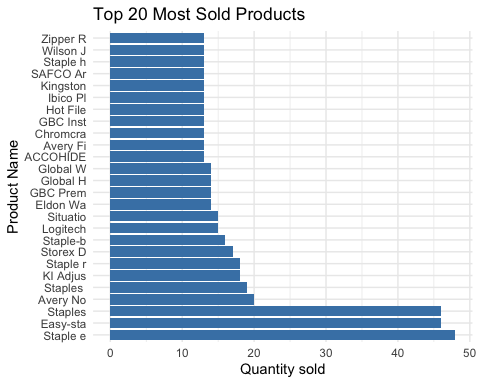
We were fascinated by the orders with a negative profit. These all had high Discount rates, and often concerned the same items, such as the Cubify CubeX 3D Printer Triple Head Print. The orders with a negative Profit where often part of a larger order (for instance CA-2016-108196), and placed by customers that placed multiple orders. We suspect these negative Profit’s to be caused by faulty items that receive discounts, general discount codes, or volumne 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 whatshowever, we suspect the chance of some weird but valid orders to be higher than encountering mistakes here. *[this paragraph could use some rewriting]*

## 2 Data Visualization

# A tibble: 20 × 3  
 Product\_Name total\_quantity ProdName8  
 <chr> <dbl> <chr>   
 1 Staples 215 "Staples"  
 2 Staple envelope 170 "Staple …  
 3 Easy-staple paper 150 "Easy-st…  
 4 Staples in misc. colors 86 "Staples…  
 5 KI Adjustable-Height Table 74 "KI Adju…  
 6 Avery Non-Stick Binders 71 "Avery N…  
 7 Storex Dura Pro Binders 71 "Storex …  
 8 GBC Premium Transparent Covers with Diagonal Lined … 67 "GBC Pre…  
 9 Situations Contoured Folding Chairs, 4/Set 64 "Situati…  
10 Staple-based wall hangings 62 "Staple-…  
11 Chromcraft Round Conference Tables 61 "Chromcr…  
12 Eldon Wave Desk Accessories 61 "Eldon W…  
13 Staple remover 61 "Staple …  
14 Global Wood Trimmed Manager's Task Chair, Khaki 59 "Global …  
15 Wilson Jones Turn Tabs Binder Tool for Ring Binders 59 "Wilson …  
16 Kingston Digital DataTraveler 16GB USB 2.0 57 "Kingsto…  
17 Fellowes Officeware Wire Shelving 55 "Fellowe…  
18 Global High-Back Leather Tilter, Burgundy 54 "Global …  
19 SAFCO Arco Folding Chair 53 "SAFCO A…  
20 Xerox 226 53 "Xerox 2…







Source: [Article Notebook](https://SJbrou.github.io/Supply_Chain_Data_Analysis/index.qmd.html)

This is a simple placeholder for the manuscript’s main document (**knuth84?**).

1 + 1

[1] 2

Source: [Article Notebook](https://SJbrou.github.io/Supply_Chain_Data_Analysis/index.qmd.html)

## 3 Introduction

Source: [Article Notebook](https://SJbrou.github.io/Supply_Chain_Data_Analysis/index.qmd.html)

|  |
| --- |
| Figure 1: Timeline of recent earthquakes on La Palma |

Source: [Article Notebook](https://SJbrou.github.io/Supply_Chain_Data_Analysis/index.qmd.html)

Source: [Article Notebook](https://SJbrou.github.io/Supply_Chain_Data_Analysis/index.qmd.html)

Based on data up to and including 1971, eruptions on La Palma happen every 79.8 years on average.

Studies of the magma systems feeding the volcano, such as Marrero et al. (2019), have proposed that there are two main magma reservoirs feeding the Cumbre Vieja volcano; one in the mantle (30-40km depth) which charges and in turn feeds a shallower crustal reservoir (10-20km depth).

Eight eruptions have been recorded since the late 1400s ([Figure 1](#fig-timeline)).

Data and methods are discussed in [Section 4](#sec-data-methods).

Let denote the number of eruptions in a year. Then, can be modeled by a Poisson distribution

where is the rate of eruptions per year. Using [Equation 1](#eq-poisson), the probability of an eruption in the next years can be calculated.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Recent historic eruptions on La Palma   | Name | Year | | --- | --- | | Current | 2021 | | Teneguía | 1971 | | Nambroque | 1949 | | El Charco | 1712 | | Volcán San Antonio | 1677 | | Volcán San Martin | 1646 | | Tajuya near El Paso | 1585 | | Montaña Quemada | 1492 | |

[Table 1](#tbl-history) summarises the eruptions recorded since the colonization of the islands by Europeans in the late 1400s.

|  |
| --- |
| Figure 2: Map of La Palma |

La Palma is one of the west most islands in the Volcanic Archipelago of the Canary Islands ([Figure 2](#fig-map)).

## 4 Data & Methods

## 5 Conclusion

## References

Marrero, José, Alicia García, Manuel Berrocoso, Ángeles Llinares, Antonio Rodríguez-Losada, and R. Ortiz. 2019. “Strategies for the Development of Volcanic Hazard Maps in Monogenetic Volcanic Fields: The Example of La Palma (Canary Islands).” *Journal of Applied Volcanology* 8 (July). <https://doi.org/10.1186/s13617-019-0085-5>.