Supply Chain Data Analytics

Analyzing and Forcasting Supermarket Sales

2024-12-05

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] "No missing values"

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

We also ran some descriptive statistics to check unlikely or impossible values, outliers, means, etc.

# A tibble: 5 × 19  
 Order\_ID Order\_Date Ship\_Date Ship\_Mode Customer\_ID Customer\_Name Segment  
 <chr> <date> <date> <chr> <chr> <chr> <chr>   
1 CA-2016-152… 2016-11-08 2016-11-11 Second C… CG-12520 Claire Gute Consum…  
2 CA-2016-152… 2016-11-08 2016-11-11 Second C… CG-12520 Claire Gute Consum…  
3 CA-2016-138… 2016-06-12 2016-06-16 Second C… DV-13045 Darrin Van H… Corpor…  
4 US-2015-108… 2015-10-11 2015-10-18 Standard… SO-20335 Sean O'Donne… Consum…  
5 US-2015-108… 2015-10-11 2015-10-18 Standard… SO-20335 Sean O'Donne… Consum…  
# ℹ 12 more variables: City <chr>, State <chr>, Postal\_Code <dbl>,  
# Region <chr>, Product\_ID <chr>, Category <chr>, `Sub-Category` <chr>,  
# Product\_Name <chr>, Sales <dbl>, Quantity <dbl>, Discount <dbl>,  
# Profit <dbl>

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

There is some more processing to do, such as removing outliers. However, by doing so we impose our own assumptions on the data (possibly the outliers are actual sales?). We will visualize and qualitatively evaluate the data first, and then decide what other processing steps to take.

First 5 Rows of the Data

| Order\_ID | Order\_Date | Ship\_Date | Ship\_Mode | Customer\_ID | Customer\_Name | Segment | City | State | Postal\_Code | Region | Product\_ID | Category | Sub-Category | Product\_Name | Sales | Quantity | Discount | Profit |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CA-2016-152156 | 2016-11-08 | 2016-11-11 | Second Class | CG-12520 | Claire Gute | Consumer | Henderson | Kentucky | 42420 | South | FUR-BO-10001798 | Furniture | Bookcases | Bush Somerset Collection Bookcase | 261.9600 | 2 | 0.00 | 41.9136 |
| CA-2016-152156 | 2016-11-08 | 2016-11-11 | Second Class | CG-12520 | Claire Gute | Consumer | Henderson | Kentucky | 42420 | South | FUR-CH-10000454 | Furniture | Chairs | Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back | 731.9400 | 3 | 0.00 | 219.5820 |
| CA-2016-138688 | 2016-06-12 | 2016-06-16 | Second Class | DV-13045 | Darrin Van Huff | Corporate | Los Angeles | California | 90036 | West | OFF-LA-10000240 | Office Supplies | Labels | Self-Adhesive Address Labels for Typewriters by Universal | 14.6200 | 2 | 0.00 | 6.8714 |
| US-2015-108966 | 2015-10-11 | 2015-10-18 | Standard Class | SO-20335 | Sean O’Donnell | Consumer | Fort Lauderdale | Florida | 33311 | South | FUR-TA-10000577 | Furniture | Tables | Bretford CR4500 Series Slim Rectangular Table | 957.5775 | 5 | 0.45 | -383.0310 |
| US-2015-108966 | 2015-10-11 | 2015-10-18 | Standard Class | SO-20335 | Sean O’Donnell | Consumer | Fort Lauderdale | Florida | 33311 | South | OFF-ST-10000760 | Office Supplies | Storage | Eldon Fold ’N Roll Cart System | 22.3680 | 2 | 0.20 | 2.5164 |

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

## 2 Section

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)).

|  |
| --- |
| Figure 3: Locations of earthquakes on La Palma since 2017 |

Source: [Explore Earthquakes](https://SJbrou.github.io/Supply_Chain_Data_Analysis/notebooks/explore-earthquakes-preview.html#cell-fig-spatial-plot)

[Figure 3](#fig-spatial-plot) shows the location of recent Earthquakes on La Palma.

## 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>.