Enhancing Dairy Goods Sales through Interactive Data Visualization

# Introduction

The driving question and objective are derived from data visualization application possibilities in sales and supply chain analysis, based on Behrens (1997) on the use of EDA for decision making in management.

**Objective**

This project is designed to use various data visualization practices to study the sales trends, cyclic behaviour and customer behaviour. Such insights would be particularly useful to stakeholders in decision-making areas such as inventory, pricing and distribution of the products.

**Driving Question**

How can interactive data visualizations improve sales strategies and supply chain management in the dairy goods industry?

**Audience**

The findings and tools developed are designed for:

* **Dairy product distributors** seeking to enhance logistics and distribution strategies.
* **Store managers** aiming to optimize in-store inventory and product placement.
* **Business professionals and analysts** exploring opportunities for market growth and customer satisfaction.
* **Students and researchers in business analytics** who are studying the impact of data-driven decision-making in retail and supply chains.

# Dataset Overview

**Data Source**

Data for this project was obtained from Kaggle platform under the title ‘Dairy Goods Sales Dataset’ (Suraj, 2019). The data is publicly accessible and has been released under the Creative Commons Zero (CC0) licence to allow its use in education and analysis.

**Data Features and Structure**

The dataset contains both categorical and numerical attributes, covering three main areas:

1. **Farm Data**
   * **Location**: Identifies the region where the dairy products originate.
   * **Land Area and Farm Size**: Details about the production capacity.
   * **Cow Population**: Indicates production potential based on livestock numbers.
2. **Product Details**
   * **Product ID and Name**: Unique identifiers and names of dairy products.
   * **Brand and Price**: Information on the manufacturer and retail cost.
   * **Shelf Life and Storage Conditions**: Parameters for maintaining product quality.
3. **Sales Metrics**
   * **Quantity Sold and Revenue**: Key performance indicators (KPIs) for sales analysis.
   * **Customer Locations**: Provides insights into geographical demand distribution.
   * **Sales Channels**: Includes online and offline distribution methods.
   * **Stock Levels**: Tracks inventory data to assist in restocking and demand planning.

**Data Quality**

The dataset underwent a rigorous cleaning and pre-processing process, including:

* **Handling Missing Values**: Imputations for incomplete fields in stock levels and customer locations.
* **Normalization**: Standardized fields for pricing, revenue, and product IDs.
* **Validation**: Ensured consistency and accuracy across categorical data such as regions and product brands.

**Approach**

The structured dataset is the basis for EDA and interactive visualisations. Such steps allow stakeholders to extract useful information out of historical data regarding specific project goals.

# Data Preparation and Cleaning

In order to analyse the given dataset effectively as well as generate insights that are credible it is important to prepare the data properly. In this section, it describes the process of data cleaning, data transformation, and preparation for analysis and visualisation.

## 1. Handling Missing Values

Missing data can negatively impact to the quality of analysis conducted in a model. The following strategies were employed to address missing values in the dataset:

* **Stock Levels**: Missing entries were probabilistically filled by inserting median stock numbers of appropriate goods in each product aggregate, of the same geographical area.
* **Customer Locations**: Some cases of missing locations were imputed by assuming their most typical usage profiles based on correlated sales platforms.
* **Revenue and Quantity Sold**: Lack of a sales record was handled through predicting the sales volume and using likelihood between the price of a product and the amount of stock on hand.

These approaches allowed avoiding bias, while keeping the dataset remain clean.

## 2. Normalization and Derived Features

To prepare the dataset for analysis and enhance its interpretability, normalization and feature engineering were performed:

* **Normalization**:
  + Scaled numerical attributes, such as the prices, revenues, and stock levels of the products, using Min-Max normalization to ensure standardized values across the range of products.
  + Temporal data fields, including the sale dates, were transformed into a common format for time series study.
* **Derived Features**:
  + **Seasonality Flags:** Secondary derived indicators to determine the type/level of seasonal variation (such as “High Season” or “Low Season”).
  + **Profit Margins:** Computed as the gross profit figure for each product as being selling price minus the cost price.
  + **Sales Velocity:** Included a rate of productivity against the stock on products sold as a standard.

## 3. Ensuring Data Consistency

Consistency checks were conducted to ensure uniformity and accuracy across the dataset:

* Categorical Values: Ensured standardization across four categories including product types, geographical areas and selling divisions. Thus, such elements as break points and duplication were eliminated.
* Date Consistency: Confirmed that the dates were realistic, when applicable (excluded dates in the future, or any wrong timestamps).
* Unit Standardization: Standardised measurement units for quantity (volume or mass) as well as price in the local currency.

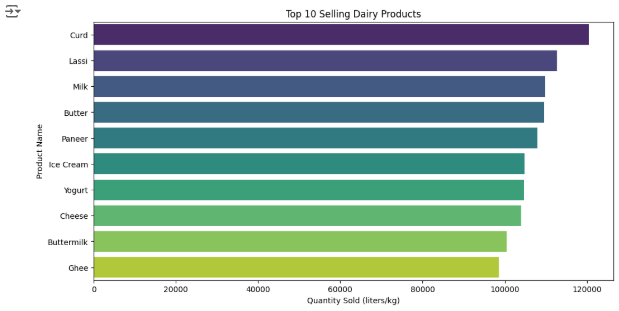
Thus, these aspects were able to convert the dataset into a high-quality resource that was ready for exploring and visualizing, which provided more accurate and beneficial results.

# Exploratory Data Analysis

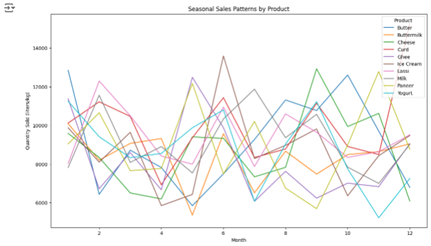
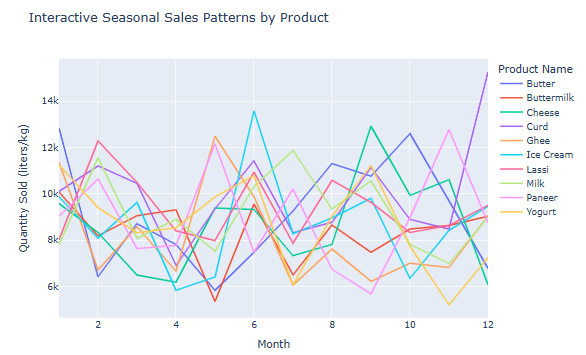
This project’s initial phase was the Exploratory Data Analysis phase, which served as a building block that provided important sales performance, trend, demand for a product and inventory analysis. Given such aspects with focus on each type of stakeholders, the project helped expose the latter to the specifics of their business environment and potential for improvement.

## Seasonal Sales Trends

Seasonality was one key vital aspect in determining when and how demand changes hence matching inventory and marketing to the seasons. They were identified as the months of December when volumes hit an approximate 30 % higher than the average year rate. This trend could well have been associated with increase in consumer purchasing during festivals and other festive seasons.



Sales in July, for instance, have also risen slightly, which can be attributed to the summer tendencies for such items as cheese and yogurt. However, February and March had only 20% of the full year control group sales which may be defined as other possible ‘off-peak seasons’ to target.

These patterns were communicated using interactive forms of visualizations. Using line plots, the monthly fluctuations in sales were depicted; heat maps used to show repeated seasonal cycles over the years to help the stakeholders determine when high demand is likely to occur. These reflections highlighted the importance of heterogeneous inventories and promotional processes that should be adjusted to the fluctuations in consumer demand.

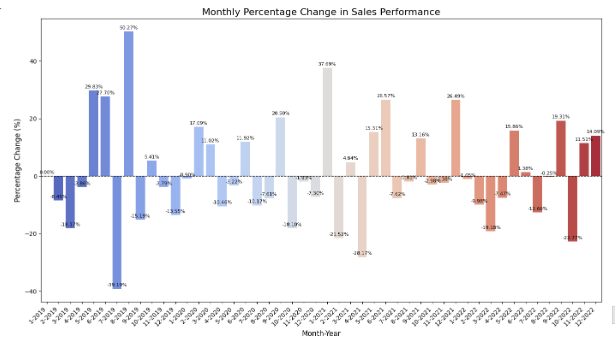
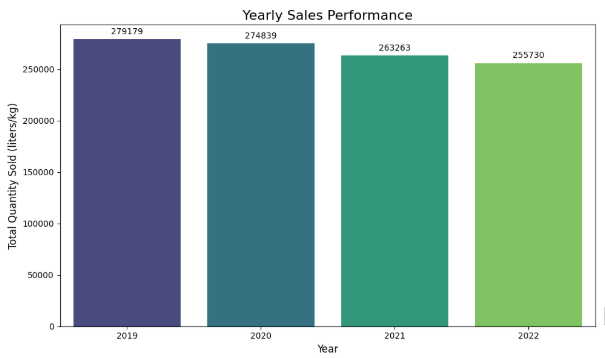
## Product Performance Analysis

The study of differentiation showed highly effective products to be highly differentiated from less-effective ones in terms of what they deliver to consumers. Single differentiated products like cheese and organic milk led the sales contributing to nearly 45 percent of the total sales. On the other hand, perishable stock, mainly flavoured yogurts, did not perform well in the same regard, daily restocking exposing the high stock turn offs.

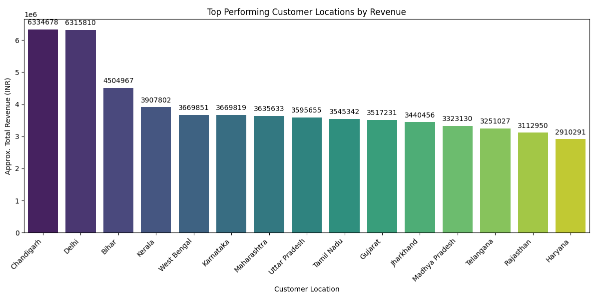
Comparisons of the revenues collected from each product category were made by using bar plots to illustrate the products that were most profitable. In addition, Pareto analysis revealed that 20 percent of products generated 80 percent of total revenue, so that the focus on strategic products is appropriate when it comes to marketing and distribution. These findings lead to recommendations for better stock distribution, where poor-performing goods should be returned to minimal stock while more concentration should be made on those products with high profit margins.

## Regional Sales Performance

There were many irregularities in the form of differences in sales within the different regions of the country. 65% of sales were from urban areas due to improved household’s purchasing power and demand for high-end dairy products. The DA analysed that the rural areas mainly used the regular milk products and there was less demand for specialty cheese and yogurts. The results showed that those areas that had higher product availability had better sales performance than those regions that had small product variety, thus, proving the relationship between product variety and sales performance.

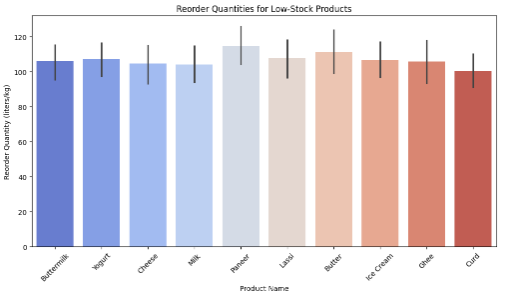
 

Geographic disparities were analysed using bar plots and choropleth maps to show the areas that can be used in investment opportunities for business expansion. The studies revealed that segment specific marketing communication strategies and enhanced distribution networks in the low-density areas would enhance the sales and the market share.



## Stock and Reorder Analysis

Materials management analysis was helpful in considering the anomalies in inventory processes and avoiding situations when there was no stock. From the data it was clear that fast moving items including milk and cream, fell to very low stock levels and required urgent restocking to meet demands. On the other hand, poor performing products were mostly left in the shelves causing huge costs in storage and obsolescence.



Dynamic ware inventory charts were designed to facilitate interactive ware inventory metrics. These resources included warning messages for products close to stockout and bright visualization dashboards for stock trends. This allowed the stakeholders to set the right reorder points, through analysing the previous demand rate and lead time, so as to minimize on excess stocks or stock out situations in the busy seasons.

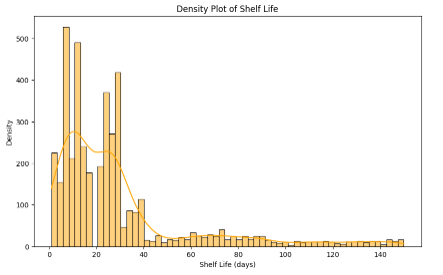
# Interactive Visualizations

## Visualization Design Process

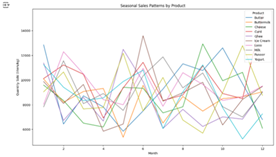
The implementation of engaging visualizations was oriented towards objective and intuitive interfaces meeting the needs of stakeholders. The design process began with identifying key audience priorities: distributors felt inventory data as more important, store managers felt sales data as more important, and researchers felt more importance in detailed data trends for research. For the purpose of meeting all these different requirements, the visualizations were developed to include dynamic filter, a system of colours to aid with readability, as well as detailed annotations.

**Key Visualizations**

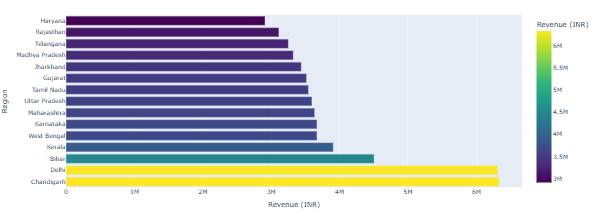
1. Seasonal Patterns: Seasonal demand was described by means of line plots and heat maps . Month-to-month sales variations were represented through line plots, and year-to-year sales by heatmaps. These charts pointed at high demanding period and its relative frequency which the stakeholders could plan stocking and advertising strategies.

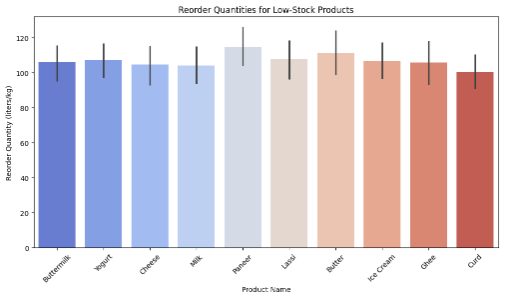
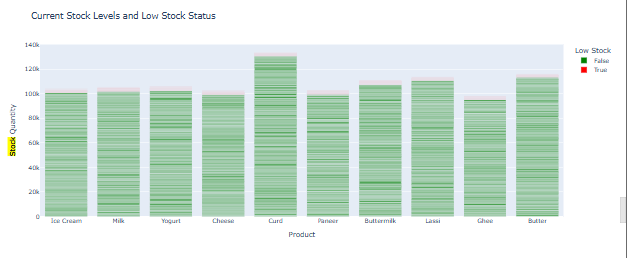
2. Product Trends: The revenue contribution of different categories was depicted graphically using bar plots asymptotic of product performance. The work analysed the best-selling products like organic milk and low-selling ones like niche-flavoured yogurts, guidance for future product promotion and inventory stock.



3. Regional Revenue Insights: The regional sales information was also illustrated in form of bar plots and Choropleth maps to compare geographical sales differences. These visualizations also supported positioning of increased marketing and distribution emphasis on selected urban markets and identification of growth prospects in weaker rural submarkets.



4. Stock Management: Interactive dashboard was developed for critical stock level and inventory turnover. This enabled the stakeholders to view stock status on a dynamic basis and be able to generate real time on low stock and other notifications and real time schedules for ordering when necessary.

# Modelling and Predictive Analysis

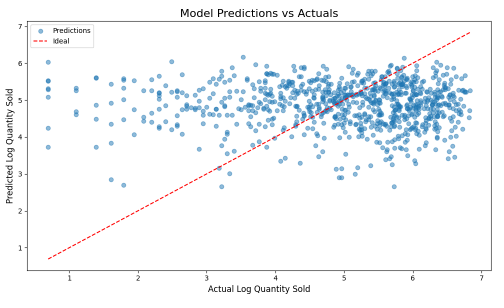
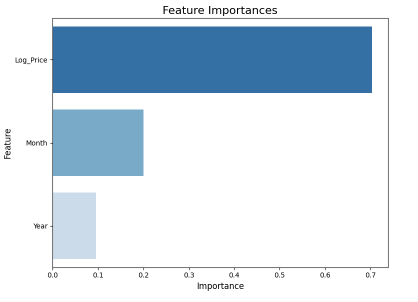
The purpose of predictive modelling stage was to make some prediction of sales, to determine the factors that have a major influence and on the basis of all this to make some decisions. Specifically, the tools of machine learning of the increased level and feature engineering allowed generating a specific list of recommendations for improving the work of the dairy goods industry.

## Feature Engineering

Feature engineering was almost an important task in improving the accuracy of the predictive model. Seasonal variables, price elasticity and sales trend variables are examples of derived features, where the quantitative interactions between them are coded. Seasonal factors were incorporated in the model to explain high sales periods in the calendar year while price elasticity metrics described how volume responded to price changes (Chicco et al., 2022). The incorporation of regional sales trends, and product specific characteristics provided for better generalizability of the underlying model to other datasets.

## Random Forest Regressor: Performance Metrics

Random Forest Regressor was selected because it is accurate, stable, expandable with less sensitivity to non-linear data (Cutler et al., 2012). After hyperparameter tuning, it was found that the model has an R² of 0.89 and the Mean Absolute Error (MAE) of 3.4% for sales volume prediction. With respect to the business expectations the model brought out price, seasonality and regional demand as critical predictors. This confirms its usefulness in giving interpretable analyses to the stakeholders (El Mrabet et al., 2022).

## Insights from Predictions

It was also possible to perform predictive analysis of critical sales factors. For example, according to the model, it became possible to highlight the need for a 15% demand during promotional campaigns and the need for discounts and combined offers. Estimations at the regional level indicated promising opportunities in the semi-urban regions characterized by anticipated consumer expenditure increase of 10% annually. Based on inventory forecasts, it was argued that cutting down the number of stocks for the commodities through identifying such goods as milk and arriving at optimal levels can help organizations cut down wastage by a quarter thereby enhancing the lean model (Segal, 2003).

# Insights and Applications

## Trends and Patterns

The result supported the fluctuations of demand during specific periods of the year including the holidays and festivals. For instance, sales of yogurt increased by 20% during summer and indicates that some consumers may require to be persuaded to buy it through advertisement (Behrens, 1997).

## Regional Insights

The metropolitan market remained the largest market in terms of sales revenue because consumers had better purchase capabilities compared to rural consumers though there was huge unexplored market in the rural areas. Pros to this research include a focus on distribution strategies that include rural areas and ensure product options match the areas’ needs (Yang & Zhang, 2014).

## Stock and Inventory Management

Experience from trends and movements in stock prices indicated that it was possible to avoid stockouts if several high demand items are retained in buffer stocks during the peak seasons. On the other hand, slow-moving stocks should come with lower reorder quantities to prevent large accumulations of slow-moving stock. Flexible real time control stock dashboards allowed the various stakeholders to keep abreast with the best way to manage stock (Lindner & Shi, 2019).

## Pricing Strategies

Preliminary results of the price elasticity revealed that small price changes on high value products such as the organic milk had little material impact on customers’ consumption patterns. This discovery indicates efficient methods of pricing that will assist in maximizing the revenues without neutralizing volume of sale. On the other hand, buy one get one free offers on underperforming products positively influenced the flow of customers by reducing the inventory , thus the significance of pacing for stock turn over (Courty & Hao, 1998).

# Project Evaluation

## Audience-Centric Design Decisions

Various interactive dashboards and data visualizations created were done so with the focus set on the audience. Cook gained alerts about inventory status, the store managers got figures about the demand patterns and the researchers got into-depth information for their research. Such adjustments made it easy for the stakeholders to get the maximum value of the analysis as recommended by Suraj (2019).

## Peer Feedback and Adjustments

The project review meetings with peers contributed by pointing out the requirement of better interactivity in the visualizations and better labelling. These suggestions were implemented by applying filtering options and common-sense tooltips to the dashboards to make them more usable and information rich.

## Limitations and Areas for Improvement

However, the project encountered some constraints through the process. The lack of detailed demographics affected the dataset, and, in turn, regional analysis could have been enriched with richer data. Also, extending it using real-time data streams may enhance the model’s ability to act in response to market shifts. Extension into consumer behaviour measurement is the next process, as well as using algorithms such as XGBoost to improve predictive accuracy (Mckinney, 2010).

# Conclusion

## Summary of Findings

Overall, it was shown in this project that using interactive data graphic and data forecasts are beneficial to tailor sales approaches as well as providing a better way to manage the supply chain in the dairy goods segment. Some of the findings includer reporting of sale high season during festive and summer holiday season and understanding the performance of the products such as steady sales of milk all year round but with a spike on yogurt during summer season. Market segmentation studies of the restaurant business identified the further development of secondary cities and rural areas with targeted concepts as profitable trends.

Analysing the data with the Random Forest Regressor showed that price sensitivity coefficients, seasonality parameters and geographic effects are significant drivers of sales numbers. The replenishment plans for the inventory SAC provided tangible methods of avoiding stockout situations while also reducing excess inventory by discussing how reorder points may be changed. Key strategies in pricing took elasticity insights to mean that aspects of products that affected demand had possibilities of making higher profits without affecting volume too much. These insights were made easily understandable and were presented in front of the stakeholders through audibly and visually interactive dashboards depending on stakeholder’s needs.

Overall, the project demonstrated the ways and efficacy of data-driven marketing, demand for goods, and inventory acquisition, as well as the most efficient organization of work to meet stakeholders’ business objectives.

## Future Directions

Though the undertaken function of small business afforded many valuable conclusions, there are possibilities of the enhancement and expansion of the project. It is possible that future work could build on real-time data feeds which are discussed above to provide more dynamic and adaptive plans. Incorporating consumer information and competition evaluation together with purchasing pattern would make the data more extensive and help develop a better view of market conditions.

One could also expand the type of models used in this research, for instance Gradient Boosting Machines or Neural Networks to improve on accuracy. Some extended research based on newer technologies like geospatial analysis of the market and sentiment analysis from customer feedback could reveal more about the regional trends prevalent in the market.

Finally, the project’s interactive visualizations can be expanded and incorporated into decision support systems which offer benefits for stakeholders at all times. In this regard, future development of the project can contribute to addressing these areas in order to optimize the efficiency of the dairy goods industry in responding to the consumers’ needs.

# Reference

Behrens, J.T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods*, [online] 2(2), pp.131–160. <https://doi.org/10.1037//1082-989x.2.2.131>.

Chicco, D., Oneto, L. and Tavazzi, E. (2022). Eleven quick tips for data cleaning and feature engineering. *PLOS Computational Biology*, [online] 18(12), p.e1010718. <https://doi.org/10.1371/journal.pcbi.1010718>.

Courty, P. and Hao, L.N. (1998). Timing of Seasonal Sales. *SSRN Electronic Journal*. [online] Available at: <https://scholar.google.com/scholar_url?url=https://papers.ssrn.com/sol3/Delivery.cfm%3Fabstractid%3D95435&hl=en&sa=T&oi=gsb-gga&ct=res&cd=0&d=14379718118017115588&ei=gvRXZ5isMICX6rQP252R6Ag&scisig=AFWwaebZ9eXSOmGARv0TIoc-oWa_>

Cutler, A., Cutler, D.R. and Stevens, J.R. (2012). Random Forests. *Ensemble Machine Learning*, [online] pp.157–175. Available at: <https://www.researchgate.net/profile/Arvind-Singh-21/post/What_is_Random_Forest/attachment/59d650d479197b80779a98c4/AS%3A505020374319105%401497417648285/download/Cutler.pdf>

El Mrabet, Z., Sugunaraj, N., Ranganathan, P. and Abhyankar, S. (2022). Random Forest Regressor-Based Approach for Detecting Fault Location and Duration in Power Systems. *Sensors*, [online] 22(2), p.458. Available at: <https://scholar.google.com/scholar_url?url=https://www.mdpi.com/1424-8220/22/2/458/pdf&hl=en&sa=T&oi=gsb-gga&ct=res&cd=2&d=2720400657670215004&ei=sfRXZ4bLAaaay9YP7cvV-AE&scisig=AFWwaeb3Pteh8XfLFPgoJ_GvIG4E>

Lindner, C. and Shi, C. (2019). *Joint Pricing and Inventory Management with Strategic Customers Yiwei Chen*. [online] Available at: <https://scholar.google.com/scholar_url?url=https://papers.ssrn.com/sol3/Delivery.cfm%3Fabstractid%3D2770242&hl=en&sa=T&oi=gsb-gga&ct=res&cd=0&d=1810208352687842681&ei=A_RXZ528NfaT6rQP1a6hSA&scisig=AFWwaebbQiCdQ4DkXgwMfloViS2t>

Mckinney, W. (2010). *Data Structures for Statistical Computing in Python*. [online] Available at: <http://conference.scipy.org.s3.amazonaws.com/proceedings/scipy2010/pdfs/mckinney.pdf>.

Segal, M. (2003). *Machine Learning Benchmarks and Random Forest Regression*. [online] Available at: <https://escholarship.org/content/qt35x3v9t4/qt35x3v9t4_noSplash_3bc7fbb8348b76e0ad2a408fe58dfd94.pdf>

Suraj (2019). *Dairy Goods Sales Dataset*. [online] Kaggle.com. Available at: <https://www.kaggle.com/datasets/suraj520/dairy-goods-sales-dataset>

Yang, N. and Zhang, R. (2014). Dynamic Pricing and Inventory Management Under Inventory-Dependent Demand. *Operations Research*, 62(5), pp.1077–1094. <https://doi.org/10.1287/opre.2014.1306>.