

## Data Source:

The **Retail Store Inventory Forecasting Dataset** is sourced from [Kaggle](#) that provides realistic data for analyzing and forecasting retail store inventory and contain around 73100 rows of daily data across multiple stores and products, including attributes like sales, inventory levels, pricing, weather, promotions, and holidays.

The primary goal of this dataset is allowing data scientists to refine supply chain efficiency and operational performance by exploring time series forecasting and assessing the impacts of weather and holidays on sales.

## Data Collection:

The Retail Store Inventory Forecasting Dataset aggregates publicly available data from various sources, including sales, inventory levels, pricing information, weather data, promotional activities, and holiday schedules. This diverse collection provides a comprehensive foundation for analyzing retail performance and forecasting inventory needs. By integrating external factors such as weather and holidays, the dataset enables a deeper understanding of their influence on sales, resulting in more accurate predictions.

Data scientists can leverage this dataset to identify patterns, optimize supply chain operations, and enhance decision-making processes. Its rich and well-curated structure ensures applicability to real-world retail scenarios and strategic planning.

- **retail\_store\_inventory.csv:** This dataset has 15 columns and contains data of store categorized by products across regions.

## **Data Profile:**

### **Cleaning & Understanding the Data**

#### **Data Types:**

Date	object
Store ID	object
Product ID	object
Category	object
Region	object
Inventory Level	int64
Units Sold	int64
Units Ordered	int64
Demand Forecast	float64
Price	float64
Discount	Int64
Weather Condition	object
Holiday/Promotion	int64
Competitor Pricing	float64
Seasonality	object

#### **Missing Values & Duplicate Values:**

There are no missing values and no duplicate values in the dataset.

## Descriptive Statistics:

Descriptive Statistics Analysis is performed in Jupyter.

	Inventory Level	Units Sold	Units Ordered	Demand Forecast	Price	Discount	Holiday/Promotion	Competitor Pricing
count	73100.000000	73100.000000	73100.000000	73100.000000	73100.000000	73100.000000	73100.000000	73100.000000
mean	274.469877	136.464870	110.004473	141.494720	55.135108	10.009508	0.497305	55.146077
std	129.949514	108.919406	52.277448	109.254076	26.021945	7.083746	0.499996	26.191408
min	50.000000	0.000000	20.000000	-9.990000	10.000000	0.000000	0.000000	5.030000
25%	162.000000	49.000000	65.000000	53.670000	32.650000	5.000000	0.000000	32.680000
50%	273.000000	107.000000	110.000000	113.015000	55.050000	10.000000	0.000000	55.010000
75%	387.000000	203.000000	155.000000	208.052500	77.860000	15.000000	1.000000	77.820000
max	500.000000	499.000000	200.000000	518.550000	100.000000	20.000000	1.000000	104.940000

## Data Integrity:

1. Data is updated 2 months ago, so it means the dataset is recent and meets the project requirement.
2. Data is verified as structured and ready for further analysis.

## Data Limitations:

1. **Data Coverage:** The dataset may not encompass all possible variables influencing sales, such as local economic conditions, competitor activities, or unrecorded events.
2. **Bias in Data Sources:** The data reflects the characteristics of the stores and regions included, which may not generalize well to other locations or retail segments.
3. **Missing or Incomplete Data:** Publicly available datasets often have gaps, such as missing values for certain periods, products, or variables like promotional data or weather conditions.
4. **External Factors Exclusion:** While some external variables like weather and holidays may be included, others such as shifts in consumer behavior, inflation, or global supply chain disruptions may be missing.

5. **Lack of Reporting:** This dataset may lack detailed documentation about how data was collected, cleaned, or processed, which can restrict reproducibility and understanding of the dataset's quality.

## Why you've chosen this data set?

I chose this dataset because it offers a practical and comprehensive foundation for addressing real-world retail challenges. **Exploratory Data Analysis (EDA)** allows us to identify patterns, seasonal trends, and relationships between variables, enable the development of accurate forecasting models. By leveraging this dataset, I can gain valuable insights into optimizing inventory management, improving supply chain efficiency, and enhancing decision-making processes.

This dataset also aligns with my personal interests by giving me an opportunity to work on meaningful retail problems, make it an excellent choice for developing and showcasing analytical and machine learning skills.

## Ethical Considerations :

### 1. Data Privacy:

- Ensure that the dataset contains customer's information, it is important to anonymize any personally identifiable information (PII) to protect customer privacy and collected with informed consent and in accordance with data protection laws like GDPR, CCPA, or similar regulations.
- Since the data is sourced from public or open platform, it's essential to use the data in a manner consistent with the licensing terms and to acknowledge the data sources appropriately.

## 2. Data Security:

- Protect Sensitive Retail data, such as sales figures, inventory levels, and pricing strategies, can be valuable and proprietary. It's important to ensure that sensitive business data is protected from unauthorized access or misuse.

## 3. Data Bias:

- The dataset may include biases based on location, time, or socio-economic factors.
- Forecasting models should aim to provide equitable inventory distribution across stores, preventing overstocking in affluent areas and understocking in underserved regions, which could disadvantage certain groups of customers.

## 4. Transparency:

- Provide clear documentation on how data was collected, processed, and analyzed, allowing for accountability in the forecasting process.
- Ensure that forecasting models are transparent and interpretable which helps stakeholders to understand how decisions are made, especially when inventory decisions can impact the availability of products for customers.

## **Define questions to explore:**

1. What seasonal trends or patterns can be observed in sales data?
2. How do external factors like holidays or weather conditions influence inventory levels? Are there any correlations between product pricing and sales volume?
3. How do sales vary across different stores or regions? What insights can be derived from analyzing promotions and their impact on sales?

4. What forecasting techniques would we use to predict inventory levels? How do you handle the time-series nature of sales data in the model?
5. How do you communicate the limitations and uncertainties of your forecasts to stakeholders?
6. What strategies could be implemented based on the forecasting results to reduce costs in the supply chain?
7. How could forecasting help manage inventory during peak shopping seasons, such as holidays?