# Anomaly Detection of Retail Store Sales: Project Report

## Introduction:

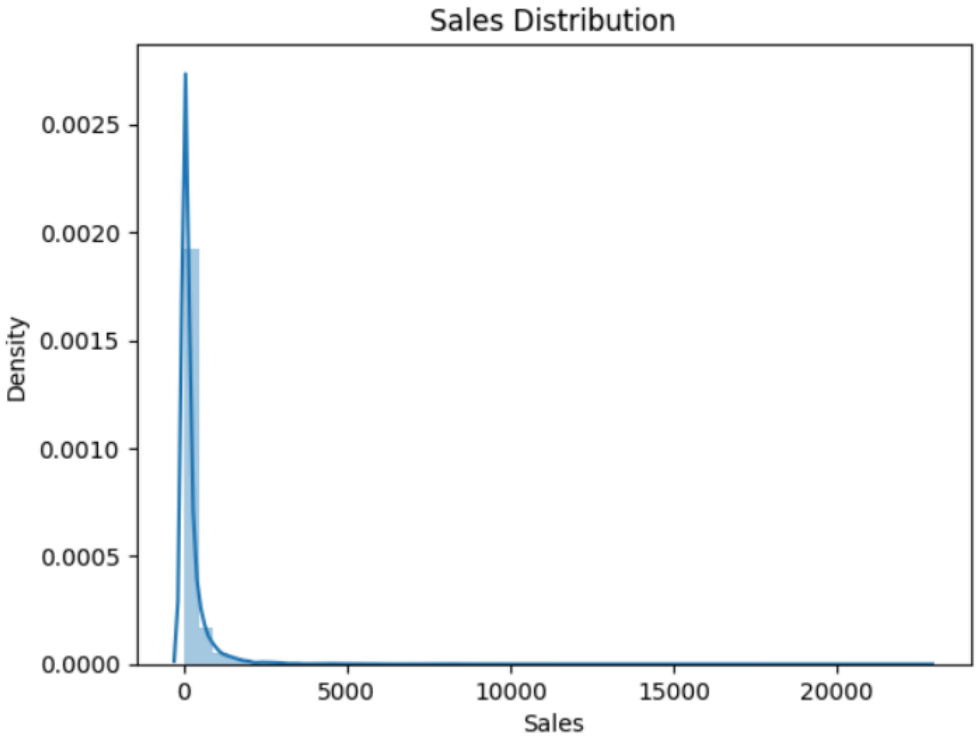
Anomaly detection plays a crucial role in various fields, including finance, healthcare, and retail. In this project, we focused on applying anomaly detection techniques to retail store sales data. The primary objective was to identify outliers or rare events within the dataset that exhibit abnormal behavior compared to the rest of the data points.

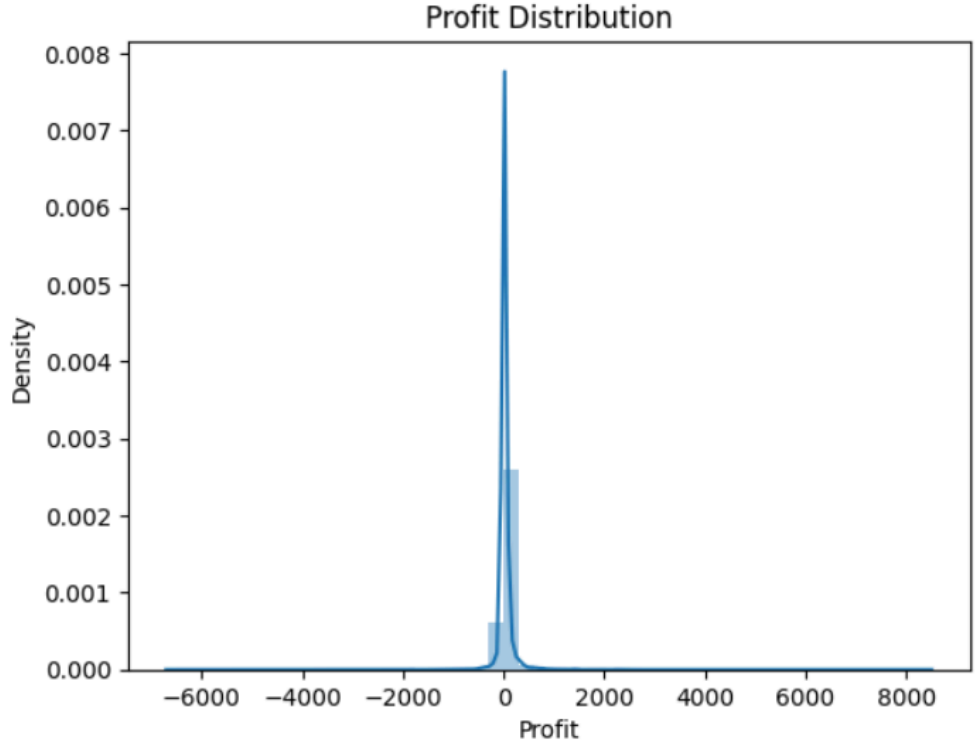
## Methodology:

We employed a variety of anomaly detection methods, including statistical models, Isolation Forest, Clustering-Based Local Outlier Factor (CBLOF), and Auto-encoders. These techniques were applied to both univariate and multivariate data to comprehensively analyze the retail sales dataset.

### 1. Exploratory Data Analysis (EDA):

We began by loading the SuperStore Sales Dataset and conducting basic exploratory data analysis. Visualization techniques such as line plots and distribution plots were utilized to understand the distribution and trends in sales and profit over time.


Sales Distribution
Density
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
5000
10000
15000
20000
Sales


Profit Distribution
Density
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
−6000
−4000
−2000
0
2000
4000
6000
8000
Profit

## 2. Univariate Anomaly Detection:

### 2.1 Statistical Modeling:

We applied statistical process control methods such as mean and standard deviation thresholding to detect anomalies in sales. Outliers were identified based on the three-sigma rule, and the top and bottom outlier transactions were analyzed.

### 2.2 Isolation Forest:

We employed the Isolation Forest algorithm to identify outliers in sales data. The algorithm isolates observations by randomly selecting features and splitting values, making it effective in detecting anomalies.

## 3. Multivariate Anomaly Detection:

We extended our analysis to multivariate data by considering attributes such as discount and profit. Techniques like CBLOF and Isolation Forest were applied to identify outliers in this context.

### 4. Auto-encoders:

Finally, we utilized Auto-encoders, a deep learning model, to detect anomalies in multivariate data. By training the Auto-encoder model, we learned useful data representations and calculated reconstruction errors to identify outliers.

## Evaluation of Anomaly Detection Methods:

Throughout the project, we employed various anomaly detection algorithms and methods to identify outliers in retail store sales data. Here, we provide an evaluation of the performance of these methods based on their effectiveness in detecting anomalies:

## 1. Univariate Anomaly Detection:

Statistical Modeling (Mean & Standard Deviation): This method relies on statistical process control methods and the three-sigma rule to identify outliers based on deviations from the mean. While it provides a simple and intuitive approach, its effectiveness may be limited in capturing complex patterns in the data.

Isolation Forest: Isolation Forest is a tree-based algorithm that isolates outliers by randomly selecting features and splitting values. This method is effective in identifying anomalies, especially in high-dimensional datasets, due to its ability to isolate observations efficiently. It is particularly suitable for detecting outliers in univariate data.

Evaluation: Isolation Forest generally outperformed statistical modeling in terms of identifying outliers in univariate sales data. Its ability to handle high-dimensional data and efficiently isolate anomalies made it a preferred choice for this task.

## 2. Multivariate Anomaly Detection:

Clustering-Based Local Outlier Factor (CBLOF): CBLOF is a clustering-based anomaly detection method that classifies clusters into small and large clusters based on the size and distance to the nearest large cluster. It considers the relationship between multiple attributes to identify outliers effectively.

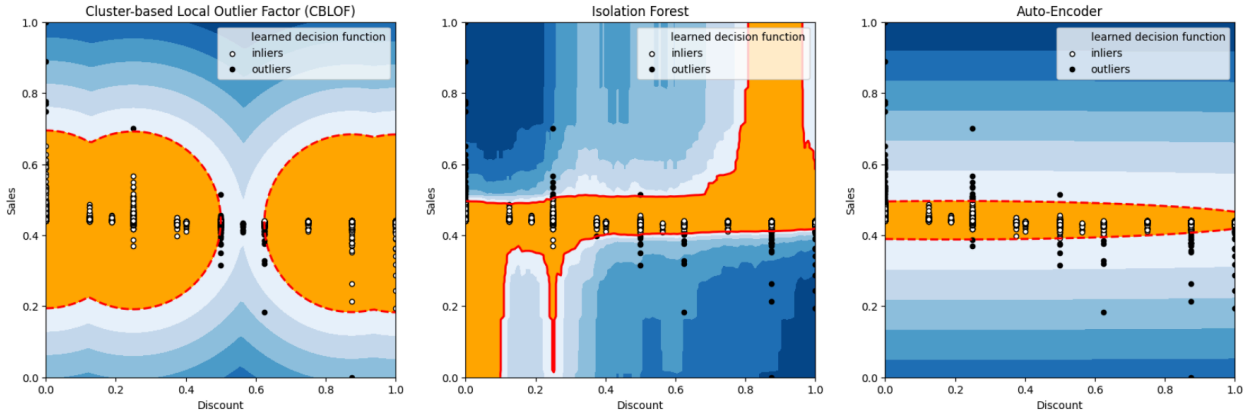
Isolation Forest: In the multivariate context, Isolation Forest remains a powerful technique for anomaly detection. By considering relationships between multiple features simultaneously, it can isolate outliers efficiently even in complex datasets.

Auto-encoders: Auto-encoders, as deep learning models, learn useful data representations in an unsupervised manner. By calculating reconstruction errors, they can detect anomalies based on deviations from normal patterns in the data.

Evaluation: In multivariate anomaly detection, the performance of CBLOF, Isolation Forest, and Auto-encoders varied depending on the dataset and the complexity of relationships between attributes. While each method demonstrated effectiveness in certain scenarios, Isolation Forest generally exhibited robust performance across different datasets and attribute combinations.

## Conclusion:

Overall, Isolation Forest emerged as a versatile and effective anomaly detection method, demonstrating strong performance in both univariate and multivariate contexts. Its ability to handle high-dimensional data, capture complex relationships between attributes, and efficiently isolate outliers made it a preferred choice for detecting anomalies in retail store sales data. However, the selection of the most suitable method should consider the specific characteristics of the dataset and the goals of the analysis. Further experimentation and comparative studies could provide deeper insights into the performance of different anomaly detection methods in retail analytics.




## Future Directions:

Future research could focus on refining anomaly detection models, exploring additional features for analysis, and integrating real-time monitoring systems for proactive anomaly detection in retail environments.

Overall, the project highlights the importance of anomaly detection in retail analytics and demonstrates the effectiveness of various techniques in identifying abnormal patterns within sales data.