## **Anomaly Detection**

**Problem Statement**

**Anomaly Detection in Sales Data**

Identify unusual sales patterns across stores and departments.

Investigate potential causes (e.g., holidays, markdowns, economic indicators).

Implement anomaly handling strategies to clean the data for further analysis.

## **What Is Anomaly?**

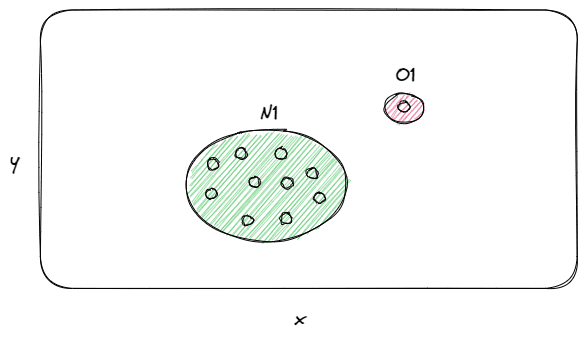
Anomalies are deviations from expected behavior or patterns in data, and their presence can indicate issues or security threats.

Handling anomalies involves a set of actions and strategies to respond effectively when they are detected.

Anomalies can occur in almost any type of data, from financial transactions and sensor readings to natural phenomena and customer behavior. These deviations can manifest in various forms and are broadly classified into three categories:

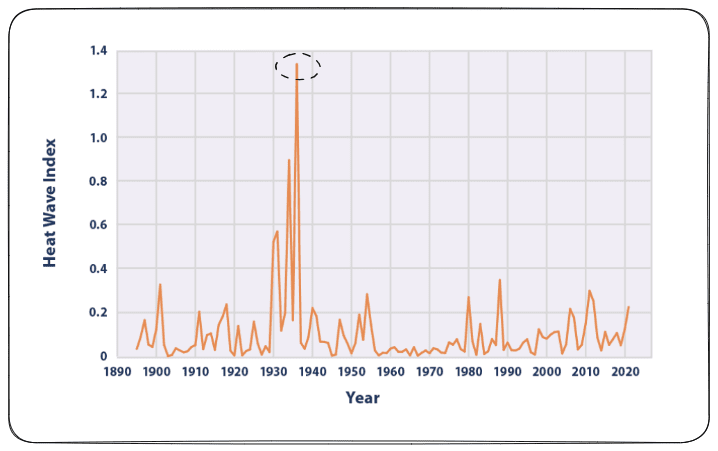
**Point Anomalies: Individual Outliers**

These are isolated data points that stand out due to their extreme values or unique characteristics. For example, a sudden surge in website traffic or an unusually high-priced item in an online store can be considered point anomalies.



**Contextual Anomalies: Context-Dependent Deviations**

Contextual anomalies are data points that are considered anomalies only within a specific context or condition. One instance to consider is the occurrence of summer heatwaves in the United States. A notable observation reveals a significant surge in the year 1930, marking an extraordinary occurrence within the country known as the Dust Bowl. This term was coined due to a prolonged period of dust storms wreaking havoc in the south-central region of the United States. ❄️🍦



**Collective Anomalies: Pattern Deviations**

Collective anomalies involve a group of data points or patterns of behavior that deviate from the expected norm. Detecting a recurring pattern of failed login attempts on a website could indicate a security breach and is an example of a collective anomaly.

### **Understanding Anomalies in Retail Data**

Anomalies in retail data refer to unexpected and irregular patterns in customer behavior or transaction data that deviate significantly from the norm. These anomalies can take various forms:

* **Unusual Purchase Patterns 🛍️📈:** Anomalies may include sudden and significant spikes in purchase orders, whether in terms of quantity or transaction amount, that do not align with historical buying behavior.
* **Outliers in Customer Behavior 🧑🏻‍💼📉:** Some customers may exhibit behaviors that deviate significantly from the norm. For example, a customer may abruptly change their buying preferences or frequency of purchases.
* **Seasonal or Event-Driven Spikes 🎉🎁:** Anomalies can occur during special events, holidays, or sales promotions when there is a sudden and substantial increase in sales or website traffic.
* **Fraudulent Activities 🕵️‍♂️💳:**Retail data may contain anomalies related to fraudulent activities, such as unauthorized transactions, stolen credit card usage, or account takeovers.

### **Why Anomalies Matter in Retail Data**

Understanding and managing anomalies in retail data are of paramount importance due to several compelling reasons:

* **Enhanced Decision-Making 📈📊** Anomalies often contain hidden insights that can inform data-driven decision-making. Identifying a sudden spike in demand for a particular product can prompt retailers to adjust inventory levels or marketing strategies to capitalize on the opportunity. 💰🚀
* **Fraud Detection 🔍🕵️‍♀️** Retail businesses are susceptible to various types of fraud, including fraudulent transactions, returns, and employee theft. Anomaly detection can help identify these irregularities and prevent financial losses. 💳🚫
* **Efficient Inventory Management 📦🏬** Anomalies in inventory data can help retailers optimize stock levels, reduce overstocking, and minimize instances of out-of-stock products. This leads to cost savings and improved customer satisfaction. 💰🛍️
* **Improved Customer Experience 😃🛍️** Understanding anomalies in customer behavior can lead to better customer experiences. For example, identifying a sudden increase in customer complaints can prompt swift action to address underlying issues and enhance customer satisfaction. 📞👥
* **Supply Chain Optimization 🚚🌐** Anomaly detection can be crucial in identifying disruptions in the supply chain, such as delays in shipments or unexpected spikes in demand. This enables retailers to take proactive measures to mitigate the impact and maintain smooth operations. 📦🔄

### **Methods for Detecting Anomalies in Retail Data**

Detecting anomalies in retail data is a multifaceted task that relies on a combination of statistical methods, machine learning algorithms, and domain expertise. Each of these methods offers a unique approach to identifying irregularities or unusual patterns in data, and they play a critical role in improving decision-making, security, and operational efficiency in the retail sector.

**Descriptive Statistics**:

Descriptive statistics are among the foundational methods for anomaly detection in retail. Techniques like mean, median, and standard deviation can be used to calculate summary statistics, which help identify point anomalies.

**Time-Series Analysis :**

Time-series analysis is particularly relevant in the retail industry, where data often involves a temporal component. Methods like Exponential Smoothing or Seasonal Decomposition of Time Series (STL) allow retailers to analyze historical sales data and identify anomalies in patterns or trends over time.

**Machine Learning :**

Machine learning has emerged as a powerful tool for anomaly detection in various retail data streams. Supervised and unsupervised techniques like Isolation Forests, One-Class SVMs, or Autoencoders can be trained to recognize anomalies in transaction data, customer behavior, or inventory levels.

**Rule-Based Systems :**

Rule-based systems offer a straightforward way for retailers to define rules or thresholds that trigger alerts when certain conditions are met. These rules can be tailored to specific business needs. For example, setting a rule to flag any transaction above a certain monetary value as suspicious or triggering an alert when there is a sudden increase in returns. Rule-based systems provide a proactive approach to anomaly detection and enable rapid response to potential issues.

**Predictive Analytics :**

Predictive analytics leverages models to forecast future behavior and compare it to actual data. Deviations from the predicted values can highlight anomalies. For instance, a predictive model might forecast the demand for a specific product in a store. If the actual sales significantly deviate from the forecasted values, it can signal an anomaly, prompting retailers to investigate the root causes, such as supply chain disruptions or changing customer preferences.

**Mitigation Strategies**

There are a number of strategies that can be used to mitigate the impact of anomalies without removing the data points. These strategies include:

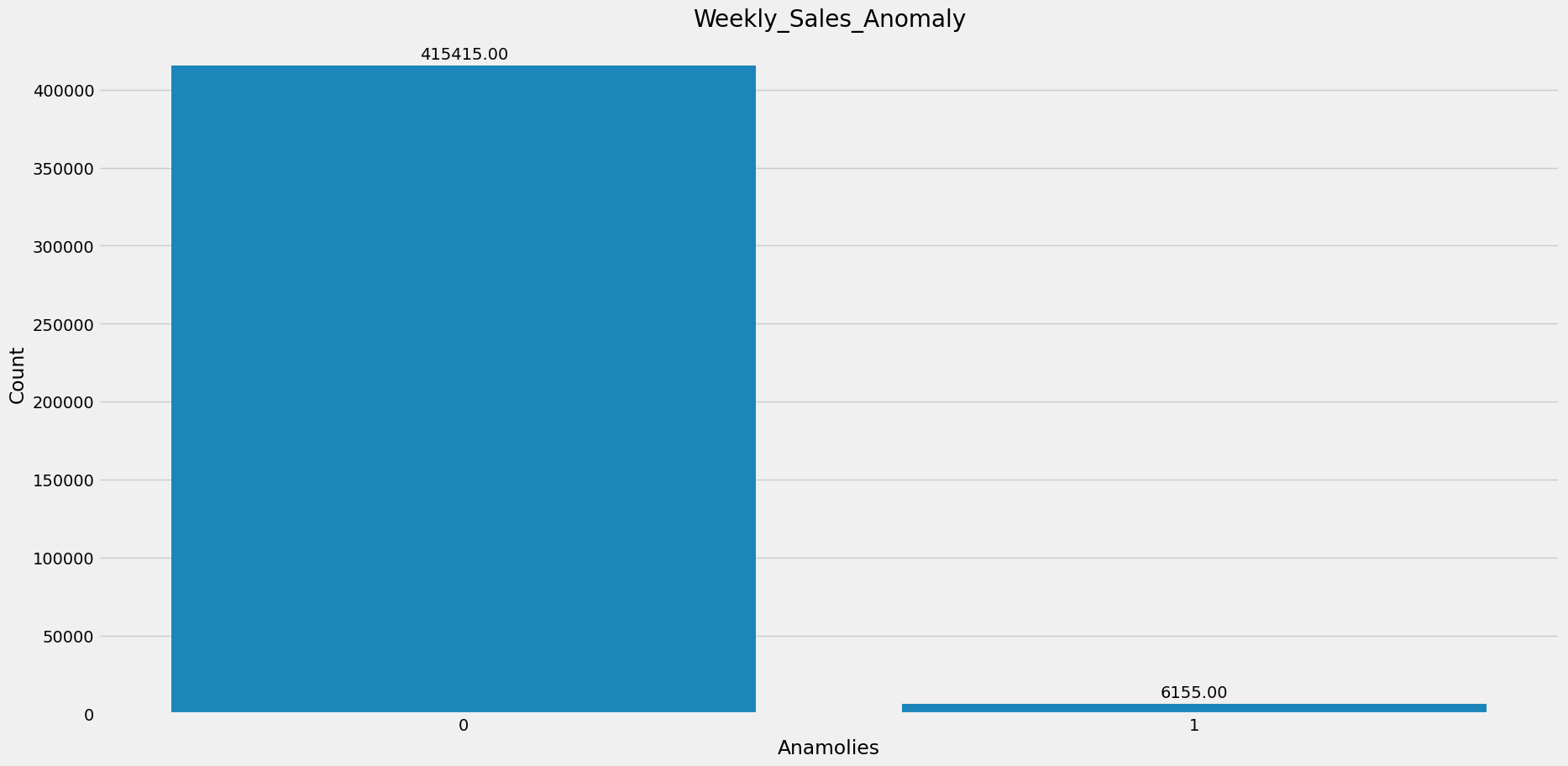
**Data cleaning:** This involves identifying and correcting errors in the data. This can be done manually or using automated tools.

**Data normalization:** This involves transforming the data so that it is distributed normally. This can help to identify anomalies more easily.

**Data imputation:** This involves filling in missing data points. This can be done using statistical methods or machine learning algorithms.

## 

## **Anomalies in Weekly Sales**



## **Anomalies in Weekly\_Sales: 6155**

**Rationale**Weekly sales data often exhibit regular patterns and fluctuations due to various factors such as seasonality, promotions, and consumer behavior. However, anomalies in this data can provide valuable insights into exceptional events or irregularities that deviate significantly from the norm. By identifying anomalies, businesses can uncover potential issues, opportunities, or areas for improvement in their sales processes.

Implement a range of techniques to detect anomalies in the dataset. This includes statistical methods like Descriptive Statistics, Z-score and IQR (Interquartile Range) as well asadvanced machine learning algorithms like Isolation Forest and One-Class SVM.

**MethodologyDescriptive Statistics:**

We begin by analyzing the distribution of weekly sales data using descriptive statistics. Descriptive statistics provide a summary of key characteristics of the data, including measures of central tendency (mean, median), dispersion (standard deviation), and shape (skewness, kurtosis).

**Standard Deviation Threshold**:

To identify anomalies, we utilize the standard deviation as a measure of dispersion. The standard deviation quantifies the average distance of data points from the mean. By setting a threshold based on a multiple of the standard deviation, we can identify data points that fall significantly outside the expected range.

**Threshold Calculation**:

In our approach, we set the threshold for anomaly detection to three times the standard deviation. This choice of threshold is based on the empirical rule, which states that approximately 99.7% of data points in a normally distributed dataset fall within three standard deviations of the mean. Therefore, data points beyond this threshold are considered potential anomalies.

**Identification of Anomalies**:

After setting the threshold, we identify data points in the weekly sales dataset that exceed the threshold value. These data points represent instances where the observed sales deviate significantly from the expected values based on historical trends.

**Implications**

**Early Detection of Issues**: By flagging anomalies in weekly sales data, businesses can proactively identify issues such as inventory discrepancies, pricing errors, or operational challenges.

**Opportunity Identification**: Anomalies in sales data may also indicate opportunities for revenue growth or optimization. For example, sudden spikes in sales may reveal the effectiveness of marketing campaigns or customer engagement initiatives.

**Data Quality Assurance**: Anomaly detection serves as a quality assurance mechanism for sales data. By validating the consistency and integrity of the data, businesses can ensure the reliability of their analytical insights and decision-making processes.

## **Importance of Post-Detection Anomaly Handling**

Handling anomalies is crucial for several reasons:

**Security:** Anomalies may indicate security breaches or attacks on a system.

**System Reliability:** Anomalies can affect system performance, stability, and reliability.

**Data Quality:** Addressing anomalies helps maintain data quality and accuracy.

**Compliance:** Some industries have regulatory requirements for handling anomalies

## **Key Goals of Anomaly Handling Strategies**

Effective anomaly handling strategies aim to achieve the following goals:

**Minimize False Positives:** Avoid unnecessary alarms or responses triggered by benign deviations.

**Maximize True Positives:** Detect and respond to real anomalies accurately and promptly.

**Reduce Response Time:** Minimize the time between anomaly detection and action, especially in critical scenarios.

**Optimize Resource Utilization:** Allocate resources efficiently to investigate and mitigate anomalies.

**Continuous Improvement:** Learn from past anomalies to enhance future anomaly detection and handling.

### **Removing Anomalous Data**

When anomalous data is detected, one option is to remove it from the dataset. This can be done in a number of ways-

**Removing Anomalous Data Using Interquartile Range (IQR):**

The Interquartile Range (IQR) is a robust measure of spread in a dataset that is less sensitive to outliers compared to measures such as the standard deviation. It is defined as the difference between the third quartile (Q3) and the first quartile (Q1). The IQR provides a measure of the spread of the middle 50% of the data.

**Benefits of Using IQR for Outlier Removal:**

**Robustness**: The IQR method is robust to extreme values and outliers, making it suitable for datasets with skewed distributions or anomalous observations.

**Preservation of Data Structure**: Unlike other methods that may significantly alter the dataset by removing outliers, the IQR method preserves the overall structure and distribution of the data while eliminating extreme values.

**Conclusion**

In conclusion, leveraging descriptive statistics and setting a threshold based on three times the standard deviation provides a robust approach to identifying anomalies in weekly sales data.

After anomaly detection, we have handled the anomaly using IQR technique on weekly sales features.

This methodology enables businesses to uncover actionable insights, mitigate risks, and enhance the overall performance of their sales operations.

**Reference**

<https://grow.almabetter.com/data-science/learn/full-stack-data-science/SPECIALISATION_TRACK/anomaly-Basics-Handling-Anomalies-sp/Concept-Colab>