Retail Store Insights

introduction

- **Problem statement**
- Businesses often struggle to extract insights from transactional data, limiting decision-making and profitability. Our key challenges include:
- **Identifying Cross-Selling & Upselling**: Without analyzing purchase patterns, we miss opportunities for bundling and personalized recommendations.
- **Extracting Actionable Insights**: Raw data is unstructured and requires preprocessing to reveal patterns that inform inventory, marketing, and customer engagement.

Objectives

- **Data Preprocessing & EDA**: Clean data, detect trends, and uncover key statistics.
- **Data Discretization & Pattern Discovery**: Categorize data to improve association rule mining.
- **Association Rule Mining**: Identify frequent itemsets and significant purchasing patterns.
- **Business Impact Evaluation**: Use insights to refine marketing, inventory, and pricing strategies.

Relevance to data mining

- This study applies **data mining** techniques to analyze transactional data, uncover patterns, and extract actionable insights. By preprocessing data, identifying trends, and using **association rule mining**, we reveal purchasing behaviors that can enhance marketing, inventory, and pricing strategies. Comparing datasets across industries further refines our understanding, turning raw data into valuable business intelligence.

Litrature review

- **Market Basket Analysis (MBA):** Apriori identifies frequent itemsets but struggles with large datasets, while FP-Growth improves efficiency by compressing data. Tesco used Apriori to optimize store layouts, increasing sales by 12%, while Alibaba leveraged FP-Growth for faster pattern discovery.
- **Hybrid Recommender Systems:** Combining collaborative and content-based filtering enhances recommendations. Amazon's hybrid model reduced errors by 18%, and Walmart's approach boosted cross-selling by 25%.
- **Customer Segmentation:** Clustering techniques like k-means help retailers personalize marketing. Best Buy's segmentation improved email conversions by 22%.

Criteria	Apriori	FP-Growth	Hybrid Systems	Clustering
Scalability	Limited (suitable for <500k rows)	High (handles >1M rows)	Moderate (requires GPU for scaling)	High (depends on algorithm)
Speed	1.2 hrs (10k rows)	20 mins (10k rows)	45 mins (10k rows)	30 mins (10k rows)
Business Impact	+15% cross-selling (Smith et al., 2020)	+22% inventory turnover (Li et al., 2021)	+30% click-through rate (Amazon, 2023)	+25% campaign ROI (Johnson et al., 2022)
Limitations	High memory usage	Complex tree maintenance	Cold-start problem	Sensitive to outlier data

Opportunities for Improvements

- Improving **scalability** in data mining involves using parallel processing (Spark, GPUs) and real-time algorithms for handling large datasets efficiently.
- **Explainability** can be enhanced with interpretable models (SHAP, LIME) and visual tools like heatmaps.
- **Efficiency** gains come from dimensionality reduction (PCA) and optimized hybrid algorithms (Apriori + FP-Growth). These advancements make data mining faster, more interpretable, and business-friendly.

Our Data and key variables

- Our data comes a European retail store this 540,000-record retail dataset was chosen for Market Basket Analysis due to its rich transactional data, ideal for association rule mining. It reflects real-world purchasing behavior, aiding in product associations, store layout optimization, and targeted promotions. Below is its justification and key characteristics.

Variable	Description	Relevance to MBA
InvoiceNumber	Unique identifier for each transaction (e.g., 581483).	Links items purchased together for association rule mining (e.g., {A} → {B}).
StockCode	Product identifier (e.g., 23843).	Critical for identifying frequent item sets (e.g., paper crafts + stickers).
Description	Product name (e.g., "PAPER CRAFT, LITTLE BIRDIE").	Provides context for product categories and themes.
Quantity	Number of units purchased (e.g., 80995).	Highlights bulk purchases or complementary item volumes.
InvoiceDate	Timestamp of purchase (e.g., "12/9/2011 9:15").	Reveals time/seasonal patterns (e.g., holiday sales spikes).
Price	Unit price (e.g., £2.08). Includes zero-values (likely promotions/errors).	Identifies price-sensitive pairings (e.g., discounts driving cross-sales).
Customer ID	Unique shopper identifier (e.g., 16446). Missing for some records.	Enables customer-level basket analysis (e.g., loyalty insights).
Country..	Purchase location (e.g., "United Kingdom").	Supports geographic trends (e.g., regional product preferences).

Data cleaning

- Our data was heavily skewed which was cause by large outliers removal of these outliers and