**AIN429 Data Mining Lab Project Proposal**

**Student Number : 2200765024 / 21993049**

**Name Surname : Velican Özkaya / Metehan Sarikaya**

**Project Title : Analyzing Market Basket Dynamics to Enhance Retail Revenue and Customer Experience**

# 1. Problem

In the business sector, basket analysis mirrors a detective uncovering frequent correlations between items. Employing data mining techniques such as association rules, ECLAT and DHP algorithms resembles discovering hidden patterns within an extensive array of shopping receipts. These strategies empower stores to strategize item placement on shelves and craft enticing deals that delight customers by offering items they enjoy together. Additional data mining tools, such as clustering similar products or predicting consumer behaviors, aid stores in comprehending their clientele better and refining their sales strategies. Within our project, our goal is to delve into sales data, identifying common product pairings to assist stores in amplifying sales of popular items. Given our 5699 unique products, we anticipate uncovering intriguing associations within our dataset.

# 2. Data

This Online Retail II data set contains all the transactions occurring for a UK-based and registered, non-store online retail between 01/12/2009 and 09/12/2011. The company mainly sells unique all-occasion gift-ware.

https://www.kaggle.com/datasets/mashlyn/online-retail-ii-uci

Attribute Information Of Dataset:

* InvoiceNo: Invoice number. Nominal. A 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation.
* StockCode: Product (item) code. Nominal. A 5-digit integral number uniquely assigned to each distinct product.
* Description: Product (item) name. Nominal.
* Quantity: The quantities of each product (item) per transaction. Numeric.
* InvoiceDate: Invice date and time. Numeric. The day and time when a transaction was generated.
* UnitPrice: Unit price. Numeric. Product price per unit in sterling (Â£).
* CustomerID: Customer number. Nominal. A 5-digit integral number uniquely assigned to each customer.
* Country: Country name. Nominal. The name of the country where a customer resides.

# 3. Methodology

There are many techniques used in data mining in the context of business and basket analysis. These are association rule, frequent itemset mining, apriori algorithm, and FP-Growth algorithm In our project, we will try to compare these basic methodologies on our project.

**Association Rules:** This method uncovers relationships between items frequently bought together. It identifies patterns like "if item A is purchased, then item B is also likely to be purchased." These rules help in understanding the co-occurrence of items in transactions.

**Frequent Itemset Mining:** This methodology involves identifying sets of items that commonly appear together in transactions. It helps determine the most frequent combinations of items purchased together.

**Apriori Algorithm:** It's a specific technique used in association rule mining. It discovers associations among items by identifying frequent itemsets and generating association rules based on these frequent itemsets.

**FP-Growth Algorithm:** Similar to Apriori, this algorithm is used to discover frequent itemsets more efficiently. It constructs a compact data structure (the FP-tree) to mine frequent itemsets.

These methodologies, especially association rules and frequent itemset mining, are at the core of conducting basket analysis. They help retailers or businesses understand customer buying patterns and make data-driven decisions to enhance sales strategies and customer satisfaction.

# 4. Plan

**4.1. Data Collection and Understanding**

Data Source Identification: Kaggle

Data Collection: Since we chose our data from Kaggle and it is structured we did the data collection part already.

Data Exploration: Understand the structure and format of the data, identify missing values or inconsistencies, and perform basic exploratory analysis.

**4.2. Data Preprocessing**

Data Cleaning: Handle missing values, outliers, and inconsistencies in the dataset.

Handling missing values: Handling missing values involves dealing with the absence of data in a dataset, which can arise due to various reasons such as data corruption, human error, measuring device error, or incomplete data collection.

Normalization and standardization: Normalization and standardization are techniques used to scale numerical features in a dataset.

Encoding Categorical Variables: Our items are in string format we will apply encoding to extract information from the dataset.

Feature Engineering: Create new features if necessary, such as grouping transactions by customer, date, or product. We have ‘InvoiceDate’ feature we will convert it to the date time data type.

**4.3.Use different Data Mining Methodologies and compare them:**

Identify Frequent Itemsets: Use algorithms like Apriori or FP-Growth to discover frequent itemsets (combinations of items frequently bought together).

Generate Association Rules: Derive association rules based on support, confidence, and lift thresholds to identify meaningful relationships between items.

**4.4. Analysis and Insights Generation**

Rule Interpretation: Interpret the generated association rules to understand item associations and their significance.

Visualizations and Reports: Create visual representations (e.g., charts, graphs) and reports summarizing the discovered patterns for easier comprehension.

**4.5. Evaluation and Iteration**

Model Evaluation: Assess the effectiveness of the association rules and recommendation systems using metrics like lift, confidence, and support.

**4.6. Documentation and Presentation**

Project Documentation: Document the entire data mining process, including methodologies, findings, challenges faced, and solutions implemented.

Presentation and Communication: Prepare a comprehensive presentation highlighting the key insights, recommendations, and the value added to the business.

# 5. References

1. Chen, D. Sain, S.L., and Guo, K. (2012), Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining, Journal of Database Marketing and Customer Strategy Management, Vol. 19, No. 3, pp. 197-208.

2. Chen, D., Guo, K. and Ubakanma, G. (2015), Predicting customer profitability over time based on RFM time series, International Journal of Business Forecasting and Marketing Intelligence, Vol. 2, No. 1, pp.1-18.

3. Chen, D., Guo, K., and Li, Bo (2019), Predicting Customer Profitability Dynamically over Time: An Experimental Comparative Study, 24th Iberoamerican Congress on Pattern Recognition (CIARP 2019), Havana, Cuba, 28-31 Oct, 2019.

4. Laha Ale, Ning Zhang, Huici Wu, Dajiang Chen, and Tao Han, Online Proactive Caching in Mobile Edge Computing Using Bidirectional Deep Recurrent Neural Network, IEEE Internet of Things Journal, Vol. 6, Issue 3, pp. 5520-5530, 2019.

5. Rina Singh, Jeffrey A. Graves, Douglas A. Talbert, William Eberle, Prefix and Suffix Sequential Pattern Mining, Industrial Conference on Data Mining 2018: Advances in Data Mining. Applications and Theoretical Aspects, pp. 309-324. 2018.