FACULTY OF SCIENCE & TECHNOLOGY

MSc Data Science and Artificial Intelligence
May 2023

Implementation and Performance Evaluation of Differential Approaches in Mining Association Rules from Offline Retail Transactions

by

David Anda

Faculty of Science & Technology

Department of Computing and Informatics

Individual Masters Project

Abstract

Offline retail businesses in emerging economies currently face a challenge of meeting changing consumer demands and offering the convenience and personalised experiences offered online retail businesses. Research has shown that attempts at utilizing Association Rule Mining (ARM) to understand customer purchasing habits have resulted in high computational costs and generation of less significant results. This study aimed to investigate the extent to which incorporating differential approaches in mining association rules can improve outcomes and optimise computational resources. A transactions over a one year period from a Nigerian retail business and differential factors around time hierarchies, item value and specific selling days were used during experimentation. Findings showed that mining association rules by segmenting the dataset by days and months produced better results in terms of quality of rules compared to the traditional approach of mining rules using the entire dataset. These results indicate that incorporating these differential factors in implementations of ARM on offline retail transactions would yield positive results and improve the understanding of customer purchasing habits. Further research is needed to understand the impacts of these approaches on broader datasets and on multi-store retail setups to identify other differential approaches that could improve the association rule mining process.

Dissertation Declaration

I agree that, should the University wish to retain it for reference purposes, a copy of my dissertation may be held by Bournemouth University normally for a period of 3 academic years. I understand that once the retention period has expired my dissertation will be destroyed.

Confidentiality

I confirm that this dissertation does not contain information of a commercial or confidential nature or include personal information other than that which would normally be in the public domain unless the relevant permissions have been obtained. Any information which identifies a particular individual's religious or political beliefs, information relating to their health, ethnicity, criminal history, or sex life has been anonymised unless permission has been granted for its publication from the person to whom it relates.

Copyright

The copyright for this dissertation remains with me.

Requests for Information

I agree that this dissertation may be made available as the result of a request for information under the Freedom of Information Act.

Signed: David Anda .

Name: David Anda

Date: 26/05/2023

Programme: MSc Data Science and Artificial Intelligence

Original Work Declaration

This dissertation and the project that it is based on are my own work, except where stated, in accordance with University regulations.

Signed: David Anda .

Name: David Anda

Date: 26/05/2023

Acknowledgments

I would like to express my sincere gratitude and appreciation to the following individuals who have contributed significantly to the completion of this dissertation:

My Supervisor, Marcin Budka, for his invaluable guidance and continuous support throughout every phase of this dissertation. His insightful feedback and constructive criticism have been instrumental in shaping the direction of this dissertation.

My family and friends for their unwavering encouragement, love, and understanding throughout my master's program. Their continuous belief in me and constant support have been a source of inspiration and motivation.

All academic staff that taught me at Bournemouth University, for transferring their knowledge and providing valuable resources that have been instrumental in my study and in conducting this research.

Lastly, I would like to express my heartfelt gratitude to all the individuals who have played a role, however big or small, in the completion of this dissertation. Your support, encouragement, and contributions have made a significant impact on my academic journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

The retail industry plays a significant role in global economic growth, serving as the link between manufacturers and consumers by making goods and services available on demand. As businesses of all sizes work to ensure the availability of a wide range of products, the retail industry remains a fundamental driver of economic activity in developed and developing countries alike. The importance of the retail industry stems from several key factors such as the numerous employment opportunities it provides especially for less educated people, the fostering of innovation and entrepreneurship among individuals, and an influence on community development (Hameli 2018).

As a result of the COVID-19 pandemic, global retail sales fell by 2.9% in 2020 affecting businesses worldwide, and despite a 10% growth bounce back in 2021 and a total of 27.3 trillion U.S dollars in global retail sales in 2022, the impacts of the pandemic, cost-push as a result of disruptions to global supply chains, and low wage growth have contributed to the current cost of living crisis that continues to impact the success of retail businesses (Akram et al. 2021; Tugba Sabanoglu 2022). Nonetheless, emerging economies, particularly Africa, are witnessing growth in retail sales due to fast rates of urbanization, expanding middle-class populations, infrastructure development and a growing appetite for consumer goods (Lisa et al. 2022). However, competition in the retail sector is intense and consumer preferences are constantly changing. Traditional retail models have been disrupted by technological advancements provided by online retail platforms that offer convenience and personalized experiences that enhance customer service. As a result, brick-and-mortar retailers face challenges to remain competitive and must continue to adapt and embrace strategies to improve offline experiences and deliver on the expectations of consumers.

Depending on the size of the business, retail transactions generate vast amounts of data that can be leveraged to extract meaningful insights to optimize business operations to improve service delivery, tailor business activities, improve marketing strategies, manage inventory, and enhance customer experience. In recent years, Association rule mining (ARM) has seen widespread applications to understand customer behaviour and identify associations between items purchased by customers to improve operational efficiency. Applications of ARM have seen improved business decisions for targeted marketing campaigns, product promotions, recommendation engines and store layout improvements that have enhanced customer satisfaction and increased profitability. However, most applications have been on online store setups and involve using retail datasets in their entirety to identify these item associations which require high-performance computing resources due to the size of most datasets. This is a challenge for offline retail businesses due to the general knowledge gap and the requirement to extract data transactions separately and implement algorithms that require significant computational resources and high-end hardware due to rule mining algorithms requiring multiple scans on these datasets. Another limitation is the inability

to discover associations present only on specific subsets of the dataset which may be missed when analyzing the entire dataset. In light of this, differential approaches that involve data segmentation before mining association rules have been proposed, but there have been few implementations and several gaps in the literature that report on the evaluation of this approach.

This dissertation focuses on differential approaches to association rule mining in the context of retail transactions using a real-world Nigerian supermarket dataset of transactions recorded from 2020 to 2021. The outcome of this project would specifically be beneficial to this retail business providing this dataset as they intend to better understand their customer base and improve their marketing strategies, store layout, customer loyalty and overall profitability. It would also generally help retail businesses with limited access to technologies to efficiently uncover purchase insights from their data.

The retail industry is highly competitive, and retailers constantly strive to gain a competitive edge by understanding customer behaviour and preferences. Association rule mining provides a valuable tool for retailers to uncover hidden relationships and patterns in customer transaction data. By leveraging association rules, retailers can make informed decisions to optimize their business strategies, improve customer satisfaction, and ultimately increase profitability.

The motivation behind this research is to explore advanced techniques and methodologies for association rule mining in the retail domain. By developing novel approaches to uncovering meaningful and actionable rules, this research aims to provide retailers with valuable insights into customer purchasing behaviour, product relationships, and market trends. Such insights can inform decision-making processes and help retailers tailor their strategies to meet customer demands more effectively.

1.1 Aims and Objectives

The aim of this project is to expand on existing body of knowledge on association rule mining and develop a rule mining system to extract association rules from novel retail transactions dataset by exploring differential approaches and investigating to what extent these approaches can optimize the extraction of rules in terms of execution speed, memory consumption and interestingness of rules in comparison to traditional rule mining approaches.

The research conducted in this dissertation seeks to address the following research questions:

- **Research Question 1:** How can association rule mining techniques be effectively applied to retail transactions data to uncover interesting item patterns and associations?
- **Research Question 2:** How can differential rule mining approaches such as time hierarchies, item value and specific day sales improve the significance of association rules?

- **Research Question 3:** How can differential rule mining approaches improve the performance of association rule mining algorithms in terms of speed and memory usage?

The corresponding specific objectives of this project are as follows:

a. Collect offline retail transactions datasets from retail business and preprocess them for the generation of association rules
b. To develop an effective association rule mining system to extract association rules from retail transactions data using traditional and differential approaches
c. Conduct experiments to evaluate the performance impact of differential approaches in mining association rules in comparison to the traditional approach

1.2 Structure of the Dissertation

The remainder of this dissertation is organized as follows:

Chapter 2 provides a comprehensive review of the existing literature on association rule mining in retail transactions. It discusses various approaches, algorithms, and methodologies proposed by researchers and highlights their strengths, limitations, and applications.

Chapter 3 presents the methodology adopted in this research, including data collection, preprocessing techniques, and the association rule mining algorithms employed. It also discusses the evaluation metrics used to assess the performance of algorithms and significance of the extracted rules.

Chapter 4 presents the results of the experiments conducted using the traditional and differential approaches association rule mining approaches on a real-world retail transactions dataset.

Chapter 5 discusses the research process, details and interpretations of findings, their implications and limitations encountered.

Chapter 6 provides a critical evaluation of the objectives and research questions on this study.

Chapter 7 summarizes the research findings, discusses their implications, and presents recommendations for future research.

2 LITERATURE REVIEW

Offline retail businesses, specifically supermarkets, superstores and specialty stores generate massive amounts of transactional data through sales and are constantly looking for ways to improve business performance. This amount of data combined with a lack of resources makes it difficult to extract meaningful insights from their data to do this. ARM is one of the most common data mining techniques used to analyse transactional data to discover patterns or relationships between items in the retail industry. This technique is typically used to discover frequently occurring itemsets from a transaction dataset and to generate association rules that describe the relationships between these itemsets. The most common approach to ARM involves using the entire transaction dataset to find rules for items that meet a user-defined frequency threshold across all transactions. However, when working with large datasets, this approach has some limitations such as high computational requirements, execution speed and memory consumption. To address these issues, higher item frequency thresholds are used to reduce the search space for algorithms thereby generating rules that only reflect relationships between popular items. Despite success in some implementations, this approach results in fewer rules being generated and increases the possibility of missing niche and interesting relationships between less frequent items among transactions. To address this limitation while also optimizing for computational requirements, differential ARM has been proposed as an effective alternative approach. Nonetheless, this approach has rarely been studied and extant research has mainly focused on generating association rules using the traditional approach.

This paper will examine the effects of differential approaches, specifically in terms of time hierarchies, item contribution to revenue and specific days in generating interesting association rules and reducing computational costs. This chapter provides a background for this research and reviews the literature on traditional and differential ARM in retail. This review examines their findings and limitations, which will provide insights for the experiments conducted within this project.

2.1 Association Rule Mining

Data mining is the process of extracting hidden and potentially useful correlations, patterns, and anomalies from analyzing large amounts of data. From a retail business perspective, it is done to establish better business strategies, enhance customer satisfaction, improve inventory efficiency, and increase profits (Han et al. 2022). First introduced by Agrawal et al. (1993), ARM is a well-known data mining technique used to discover customer purchasing habits by extracting frequently occurring patterns, associations, and structures between distinct sets of items within transactional datasets. A retail transaction dataset is typically tabular structured and consists of uniquely identifiable transactions made up of a collection of different items purchased by customers on a given day and can give an overview of their current buying habits (Chang et al. 2014). Discovering, for example, that customers tend to purchase rice, bread and potatoes together can help retail

business owners improve store layouts and promote these items together. These discoveries are referred to as rules, typically expressed as "if → then" statements that describe the relationships between items in a given dataset (Solanki and Patel 2015; Telikani et al. 2020). A rule consists of two fundamental components: the antecedent (if) and the consequent (then), and each component can consist of either a single item or a collection of items referred to as itemsets. An association rule in the form of $A \rightarrow B$, where *A* and *B* are non-overlapping itemsets in a transactional dataset, indicates a customer purchasing pattern interpreted as if a customer should purchase *A* then they are likely going to purchase *B.* The subsequent section of this literature review will discuss the various types of association rules and their applications specifically within the retail domain. This will shed light on their diverse forms and practical implementations of association rules in the context of retail implementations.

2.2 Types of Association Rules

The section will first explain the concept of frequent itemsets and their significance in ARM. Before delving into the various types of association rules, it is important to understand how they merely serve as a starting point for rule mining and differ from actual rules. This will lay the foundation for comprehending the subsequent explanations of the different rule types and their applications in retail.

2.2.1 Frequent Itemsets

Frequent itemsets are items that appear frequently in a dataset. These itemsets can consist of single items or a combination of items that frequently co-occur together in a transaction dataset. 'Frequently' is determined by a user-supplied minimum threshold referred to as support, which is simply the proportion of transactions in which the itemset appears (Aggarwal et al. 2014). Frequent itemset mining is at the root of ARM and the identification of frequent itemsets serves as the foundation for generating association rules that describe the relationships between these itemsets. To generate association rules, ARM algorithms start by scanning the dataset to identify frequent itemsets above a specified minimum support threshold then it subsequently scans through these itemsets using other specified evaluation measures to extract rules that would be considered useful by the user (Koh and Ravana 2016; Luna et al. 2019). Positive association rules, derived from frequent itemsets, describe the positive correlation between the antecedent and consequent of an association and frequent itemsets are sometimes assumed to automatically translate to them. However, not all frequent itemsets imply a positive correlation between items and need to be evaluated further using other metrics to identify those that exhibit strong correlations before arriving at this conclusion. Although they seem related, they are distinct concepts in ARM and frequent itemsets only provide the starting point for rule generation, but the subsequent analysis and evaluation of the generated rules are used to identify the truly meaningful and actionable associations between items.

Originally proposed for market basket analysis in retail transactions with tabular and static transactional data (Agrawal et al. 1993), frequent itemset mining has since been extended to other retail applications where data arrives in sequences and at different periods to identify purchasing sequences that occur frequently (Fournier-Viger et al. 2017).

2.2.2 Positive Association Rules

Positive association rules, derived from frequent itemsets, are the most common type of association rules generated across all domains and can be described as the default purpose and use case for ARM (Bagui and Dhar 2019). Typical explanations and examples of ARM describe the generation of positive association rules where the antecedent represents the purchase of a set of items that lead to the purchase of other items (consequent) and all traditional ARM algorithms are designed to mine these rules. In the context of retail, these rules are useful for understanding customer habits and item dependencies from customer transactions and are applied in offline retail domains for physical store layout optimization and in-store promotions such as bundling these items together where customers are entitled to a positively associated item (consequent) for free or at a discounted price if an antecedent item is purchased, and in online setups for targeted adverts and recommendation systems which apply the same promotional strategies to suggest items that are likely to be of interest to a customer based on past purchases and site usage behaviours of other customers (M. Shridhar 2017). Common metrics used to evaluate positive association rules include support, confidence, and lift. Support represents the proportion of transactions where the rule is found in the dataset, confidence represents the accuracy of the rule by calculating the proportion of transactions containing the antecedent that also contain the consequent, and lift measures the strength of the association between both parts of the rule by comparing the observed support against the expected support if they were independent. The research problem addressed in this project focuses on mining positive association rules and a detailed critical review of the literature on positive association rule mining as well as differential approaches are presented in section 2.3.

2.2.3 Negative Association Rules

In contrast to positive association rules, negative association rules describe the relationships between the presence of antecedent itemsets with the absence of consequent itemsets, highlighting negative correlations. They identify items that are mutually exclusive and mainly occur independently in a transaction dataset (E. Bala Krishna A. Nagaraju 2015). Despite their name and the fact that they are less common, negative association rules have been proposed to be as important as positive rules in retail scenarios when devising store layout plans and marketing strategies (Duggirala and Narayana 2013; Kavitha and Selvi 2016). However, most algorithms designed for identifying negative relationships require more computational resources to generate these rules because these

item patterns are less frequent in the dataset and it is more difficult to identify meaningful patterns and establish strong associations (Mahmood et al. 2014). One of the first attempts at mining negative association rules from retail transactions was by Savasere et al. (1999) where they developed an algorithm that leaned on domain knowledge in the form of taxonomy combined with already generated positive association rules to reduce the algorithm's search space and deduce negative association rules. Another approach proposed by Wu et al. (2004), extends the traditional approach to ARM and divided the association rule mining problem into two by generating both frequent and infrequent itemsets, extracting positive rules from the frequent itemsets and negative rules from the infrequent itemsets. They evaluated their methods for rule interestingness and scalability using synthetic and real-world data and proved effectiveness and efficiency in generating both types of rules. (Kavitha et al. 2011) proposed the use of the conviction measure in conjunction with the support and confidence measures to mine positive rules and identify negative associations based on the conviction scores of those rules. However, they conducted their research on synthetic data and this cannot be generalized to be suitable for real-world applications. Negative association rule mining has not received the same level of research interest as positive association rule ming and despite the introduction of Zhang's metric (Yan et al. 2009) which measures itemsets disassociation, there is still a gap in the literature regarding its utilization and evaluation, leaving negative association rule mining relatively unexplored.

2.3 Applications of Association Rule Mining in Retail

In response to the growing demands of customers and the need to maintain competitive advantage, research has been conducted into several applications of ARM in retail, and this section examines some of these applications, describing their methodological approaches, algorithms, results obtained, challenges faced when evaluating algorithm performance and achieving intended outcomes, as well as suggesting future research areas to pursue. Due to the differences in methodology and dataset sources used in literature, this review would focus on research on market basket analysis using ARM on offline retail datasets.

To support the business operations and improve knowledge discovery of customer buying habits in a precious metal company in Indonesia, (Yudhistyra et al. 2020) used the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework to divide their solution into phases from business understanding through to model evaluation and deployment. Using a dataset provided by the company containing 3,986,872 customer transactions over a 9-year period, they performed data cleaning to remove noise and ensure the effects of duplication, incorrect and inconsistent item naming conventions did not hinder their analysis. They displayed basic item frequency and the strengths of their relationships in transactions using a web graph and then used the Apriori algorithm proposed by (Agrawal et al. 1993) on the whole dataset to generate 21 association rules with a minimum support of 1% and a minimum confidence of 50%. The strength of each rule was evaluated using the lift ratio. For performance comparison and with the hopes of finding rules with multiple

consequent items, they used the Continuous Association Rule Mining Algorithm (CARMA) algorithm (Hidber 1999) with the same minimum support and confidence values and extracted the best rules using lift. The results of their experiment showed that both algorithms produced similar product association rules, but contrary to their research on previous works, Apriori took less time to execute compared to CARMA at 107 to 120 seconds respectively. They concluded that the nature of the dataset in use determines the performance of the model. Although the authors used these results to propose changes to the company store layout and strategies for promotional campaigns, they did not explore differential and temporal effects in generating association rules considering they worked with a massive dataset spanning across several years.

In a study by Alfiqra and Khasanah (2020), they proposed an ARM approach that assumes high variability in customer purchasing behavior over a one-month period. Their dataset contained 57,784 transactions with 41,248 items. They explain the importance of computing the variability of association rules using the lift and confidence values as they show the degree of variation of the same rules in different time periods. The pair then performed data preprocessing steps to manage inconsistencies and remove unnecessary variables such as transaction date as they were irrelevant for generating rules. Transactions were split into 4 periods of weekly transactions in the month and trial and error was used for parameter setting to obtain optimal values for minimum support and minimum confidence (0.1% and 20%). They used the Apriori algorithm to generate rules for each period and combined similar rules across each period to calculate the overall variability of each association rule. Alfiqra and Khassanah then used the conclusion by Papavasileiou and Tsadiras (2011) to only filter rules with variability values greater than 30% as those rules are subject to changes over time. Their final results had 17 rules and were used to recommend marketing strategies to promote these items together and new shelf layout arrangements to put these items in close proximity. Their study experimented with generating rules over different time periods but information on the execution time did not compare the generation of rules with any other algorithm.

Nurzani and Tania (2020) aimed to find frequent item patterns to recommend promotion techniques and improve item replenishment from a retail store dataset of 58,068 transactions from a 9-month period. They focused on generating rules with multiple antecedents and one consequent and the effect it had on memory consumption and execution time. In their experiment, they compared the performance of Apriori, Frequent pattern growth (FP-Growth) and Equivalence Class Transformation (Eclat) algorithms at different minimum support and confidence values. Using the Eclat algorithm, they split their dataset into train and test sets of 80% and 20% for three different quarters and generated rules with a minimum support of 0.003 and a minimum confidence of 100% with a minimum antecedent itemset of four items. Their results showed that most rules involved the same consequent item with different naming variations as the authors did not focus on data pre-processing before mining association rules between items. Some interesting rules were still generated with high confidence and lift values. Finally, they computed a confusion matrix to evaluate the rules generated

with the actual conditions of the data in the test set and achieved an accuracy score of 88% and a recall of 92%, allowing them to confidently achieve their objectives.

2.4 Challenges and Identified Research Gap

From the above literature review, ARM is computationally challenging for most applications as it typically involves mining rules from large datasets that span across years of transactions. Another challenge is the growth and introduction of new items over time which make it more difficult to search through all possible combinations to identify significant associations. Thus, this process requires adequate computational resources to process large datasets quickly and accurately. Finally, the quality of the results can be affected by the data preprocessing steps such as handling missing values, data outliers, completeness and restructuring the dataset to meet the requirements of the algorithms. In light of these research gaps and challenges identified, the methodology, data preprocessing steps and algorithm choices would be carefully evaluated to avoid the same limitations.

3 METHODOLOGY

Due to the nature of the research problem being addressed and the aim of providing relevant answers to the research questions, the Cross-Industry Standard Process for Data Mining (CRISP-DM) was adopted to guide the overall methodological approach to achieve the project objectives. CRISP-DM is the golden standard and most adopted methodology for data mining projects and universally serves as a clear model providing a well-structured approach for solving real-world data science and analytics business problems (Schröer et al. 2021). Adopting this industry-standard methodology is thus suitable for this project as it has delivered and continues to produce positive results for much larger-scale projects.

An overview of CRISP-DM is shown in Figure 1. It consists of six sequential phases which include Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. Although these phases are approached sequentially, the overall process is iterative as understanding gained at different phases are continually used to inform actions throughout the project lifecycle. These phases were adapted to suit the requirements of this project. The Evaluation and Deployment phases as explained in the CRISP-DM methodology do not apply to the scope of this project

Figure 1: CRISP-DM Process Model (Raja 2017)

3.1 Business Understanding

This phase was used to outline and frame the research problem from a real-world business perspective. Business questions were formulated around the reasons why a differential ARM system was needed, and the high-level research questions were drilled down to be more specific. This was done to inform the differential approaches that were going to be used in the experimental phases. Some of the low-level business questions that were formed were: “Do customers purchase different sets of items during festive seasons compared to normal periods?”, “Are there specific days in the week or months in the year where customers tend to buy different sets of items?”, “Will focusing on finding only co-occurring high-value items improve our profitability?”, “Would data segmentation improve the speed of finding co-occurring items and optimize marketing campaigns?”.

The overall project implementation plan and steps that were implemented in subsequent phases were formed at this stage and the availability of data sets and tools to be used for data preparation and modeling were also verified at this stage.

3.2 Data Understanding

The datasets used were collected from the company’s IT department and were anonymized to remove personally identifiable information to protect privacy. 13 datasets in CSV file formats were collected in total. 12 of them contained monthly customer sale transactions of one of the company’s retail locations over one year from the 1st of April 2020 to the 31st of March 2021, a total of 769,083 records. The last dataset contained 39,860 attribute information of each item present in the sale transactions datasets. A customer is defined as a unique sale transaction made by a person that had purchased items on any given day over the one year. Individual customer information was not provided as most customers were not registered on the company database due to the brick-and-mortar setup of the business where customers physically browse and make purchases in person.
A transaction is represented as separate line items linked by a common identifier (ID) (Bill No), with details of each unique item such as quantity purchased (Sold Qty), item ID (Item Code), item name, sale date (Bill Dt.), purchase price and the sale price (Item Rate) of the item. The attribute dataset contained each item ID (Item Code) and its corresponding company-defined category.

The monthly transaction datasets were combined and then merged with the attribute dataset using the item ID to create a single dataset of all sale transactions with item categories for each line-item record. The goal was to avoid implementing data cleaning steps multiple times in the preparation phase and also because a single dataset was required as a base point for the modeling phase. The combined dataset was then explored and visualized to understand the frequency distribution of transactions and to check for quality issues (validity, missing data, outliers, duplicates) using data manipulation and plotting Python libraries pandas and Matplotlib on a Jupyter Notebook. Data validity issues were found on some transactions where items had purchase prices recorded as 0 and in

some instances where purchase prices were equal to or less than selling prices. Table 1 details the number of records and unique transactions for each month and the total number of unique items. Figure 2 shows a snapshot of the content of the dataset.

Table 1: Number of Records and Transactions

Month	Number of Records	Number of Unique Transactions
April 2020	44586	10999
May 2020	55594	15328
June 2020	59113	19419
July 2020	60580	19681
August 2020	61678	19911
September 2020	64072	21912
October 2020	61079	19941
November 2020	72041	24573
December 2020	77726	26084
January 2021	74600	26096
February 2021	70633	24515
March 2021	67381	24097

	Bill No	Bill Dt.	Item Code	Item Name	Units	Sold Qty	Purchase Price	Item Rate	Major Category
769078	CC192683	2020-09-27	25049	TRS BASMATI RICE 10KG	1	1.0000	17,000.0000	21,350.0000	BISCUITS & SNACKS
769079	CC193801	2020-09-29	13154	MULTI PURPOSE COMPOTE SQ/ROUND	1	1.0000	1,000.0000	1,200.0000	KITCHEN & DINNING
769080	CC193568	2020-09-29	3572	SIMILAC PRO-ADVANCE INFANT 964g	1	1.0000	19,666.6700	23,600.0000	BABY- KID & MOTHERCARE
769081	CC194541	2020-09-30	1141	AMEL SUSAN BAKING FLOUR 1KG	1	1.0000	600.0000	780.0000	FOOD CUPBOARD
769082	CC194541	2020-09-30	2197	KIDS ORGANIC PROTEIN & VITAMIN	1	1.0000	1,313.0000	1,580.0000	BABY- KID & MOTHERCARE

Figure 2: Sample of Dataset

3.3 Data Preparation

In this phase, the data quality issues discovered in the data understanding phase were handled. After discussing with the dataset provider to understand the context of some of these validity issues, some of them were fixed by calculating the purchase price as a percentage difference from the sale price, one item was completely removed from the dataset as it was not a real item sold for profit. All other price validation issues were handled first by using the latest purchase and sale prices for each item among all transactions to replace the prices for these items in older transactions. Instances where purchase prices were still inaccurate after this step or were still higher or equal to sale prices, were handled by calculating the mode markup percentage value between the purchase price and sale price across other items that belonged to the same item category and then that value was used to calculate the purchase price for these items. The sale date field of the dataset was then used to

derive day and month attributes to explore the business questions formed in the business understanding phase and to inform the differential approaches used in the modelling phase.

3.4 Modelling

The modelling phase of this methodology involves selecting modelling algorithms, generating a testing design, building and assessing the model. After a review of the literature on ARM applications, Apriori and FP-Growth algorithms were chosen to implement and run experiments in this project using the mlxtend Python package. Both algorithms were assessed using execution speed and memory usage using the inbuilt Python time module and the virtual memory module of psutil (python system and process utilities), a cross-platform library for retrieving information on running processes and system utilization in Python (Giampaolo Rodola 2023).

3.4.1 Algorithms

An overview of the algorithms used and their working principles in generating frequent itemsets and association rules are explained in this section.

3.4.1.1 Apriori Algorithm

The Apriori algorithm proposed by Agrawal et al. (1993) is the most commonly used algorithm for mining association rules. It discovers frequent itemsets from transactional datasets using an iterative bottom-up and breadth-first approach to generate candidate itemsets of increasing lengths (frequent item subsets are extended one item at a time) and continues to scan the datasets to prune candidate itemsets that do not meet the user-supplied support (item occurrence frequency) threshold and only terminates when the frequent itemsets can no longer be extended (Du et al. 2016; Kavitha and Selvi 2016). It starts by scanning the dataset to calculate the support of itemsets with 1 item and prunes items that do not meet the support threshold, it then generates candidate itemsets with 2 items using the frequent itemsets generated in the previous step as a foundation and scans the dataset again to calculate the support of these candidate itemsets, itemsets that do not meet the support threshold are pruned again and these steps are repeated until there are no more itemsets that satisfy the support threshold. After all frequent itemsets are extracted, it generates association rules consisting of antecedent and consequent itemsets with a user-defined measure to indicate the strength of each rule. Although it is widely used because of its simplicity, the dimension of a dataset can result in repeated scans that generate many candidate itemsets during the search process, making it computationally expensive (Han et al. 2022). Figure 3 and Figure 4 show an example of the rule generating process using a support threshold value of 2.

TID	List of item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Figure 3: Sample Transaction Data (Apriori and FP-Growth)

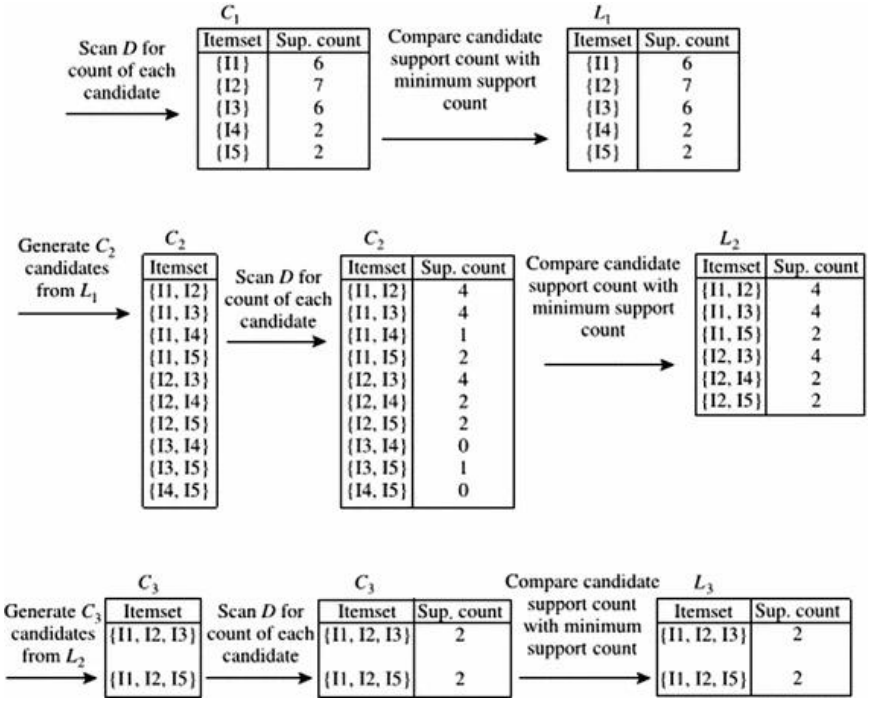


Figure 4: Apriori Frequent Itemset Generation Process

3.4.1.2 FP-Growth Algorithm

The FP-Growth Algorithm (Han et al. 2000) in contrast to the Apriori algorithm, mines frequent itemsets without generating candidate itemsets. It uses a tree-like data structure called an FP-tree (Frequent Pattern tree) to represent the frequent itemsets in a dataset depending on a user-supplied minimum support threshold value. In the first step of the rule-mining process, the dataset is scanned once to build the FP-tree using single items that meet the support threshold. In the next step, the algorithm recursively explores the FP-tree to generate frequent itemsets using a conditional pattern base technique to identify conditional FP-trees and their corresponding frequent itemsets. corresponding frequent itemsets. This advanced data compression structure eliminates the need for multiple scans, subsequently improving the speed at which frequent itemsets are found and rules are generated. This algorithm is designed to handle high-dimensional datasets with a high number of distinct items (Deng and Lv 2014). Figure 3 and 5 show the workings of this algorithm with a minimum support threshold of 2.

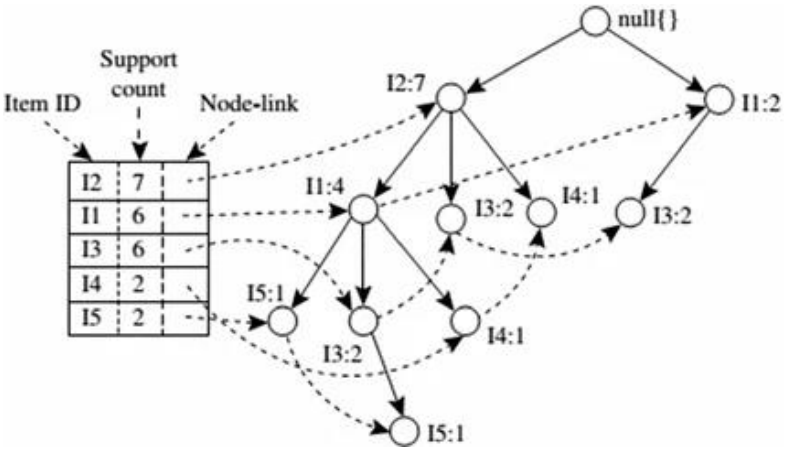


Figure 5: FP-Growth Frequent Itemset Generation Process

3.5 Evaluating Association Rules

The following metrics are currently supported in mlxtend for evaluating the quality of association rules and setting selection thresholds for extracting rules.

3.5.1 Support

Support is the relative frequency of an item in the transaction data. It expresses how popular the item is as represented by its proportion of the total items sold. It is a measure of the frequency of transactions in a database that contains an itemset. Calculated by dividing the number of transactions in the dataset that contains these items by the total number of transactions (Ali 2023). A minimum support threshold is predefined to only generate rules where itemsets have support values higher than this threshold.

$$support(A \rightarrow C) = \frac{Transactions\ containing\ both\ A\ and\ C}{Total\ transactions}$$

3.5.2 Confidence

Confidence is a measure of the reliability of a rule, which is how many times the rule actually appears among transactions. Calculated as the proportion of transactions that contain the antecedent and consequent in relation to the transactions that contain just the antecedent. Ranging from 0 to 1, where 1 indicates a positive correlation between items, it reflects the probability of the consequent item being in a transaction that the antecedent is in (Raschka 2022). The confidence of a given rule is not symmetrical as the confidence of A → C is different from C → A and can be useful to extract specific rules that focus on just one side of the rule.

$$confidence(A \rightarrow C) = \frac{support(A \rightarrow C)}{support(A)}$$

3.5.3 Lift

Lift compares the strength of the association between the antecedent and consequent items in a given rule. It measures how much more frequent items in a rule would be in the same transaction than they would be if they were independent. Calculated as the confidence of the rule divided by the support of the consequent item. The value of lift ranges from 0 to infinity, where a value of 1 indicates that the antecedent and consequent items are independent, a lift value greater than 1 indicates a positive association between the antecedent and consequent items, and a value less than 1 indicates a negative association between the antecedent and consequent items.

$$Lift(A \rightarrow C) = \frac{confidence(A \rightarrow C)}{support(C)}$$

3.5.4 Leverage

Similar to lift, leverage calculates the correlation between items in a given rule. It is a measure of the difference between the proportion of transactions where the antecedent and consequent items appear together and the proportion of transactions of their co-occurrence if they were independent. The value of leverage ranges from -1 to 1, where 0 indicates that the antecedent and consequent items are independent, leverage > 0 indicates a positive association between the antecedent and consequent items, and a negative leverage value indicates a negative association between the antecedent and consequent items meaning those items tend not to be found in the same transactions. Higher leverage values indicate a stronger deviation from independence between the antecedent and consequent items and means the rule is not due to chance.

$$Leverage(A \rightarrow C) = support(A \rightarrow C) - support(A) * support(C)$$

3.5.5 Conviction

Conviction is the measure of the dependence between the antecedent and consequent items in an association rule. It calculates the ratio of the expected frequency of the consequent item being in a transaction when the antecedent item is not to the observed frequency of the consequent item being absent when the antecedent item is present in a transaction. The value of conviction ranges from 0 to infinity, where a high conviction value (> 1) indicates the consequent is highly dependent on the antecedent, a conviction value equals 1 means the antecedent and consequent items are independent, and a value less than 1 means that the antecedent and consequent items are negatively correlated and hardly appear in the same transactions.

$$Conviction(A \rightarrow C) = \frac{1 - support(C)}{1 - confidence\ (A \rightarrow C)}$$

For the experiments conducted in this research, the lift metric was used as a rule selection criterion when generating association rules and was used to evaluate the quality of a rule (the higher the lift score, the more effective the rule is).

4 RESULTS

This chapter presents the findings and outcomes obtained from the conducted experiments in the modeling and evaluation phases of the methodology, which aimed to address the research questions and achieve the specific objectives outlined in the previous chapters. Accordingly, this chapter is subdivided into a series of sections that provide a concise and objective account of the data and performance results of the algorithms and approaches used for implementation and analysis, and how these approaches tie back to the research questions. These results only state the specific outcomes of the experiments before a comprehensive analysis and discussion are delved into in the following chapter.

First, both ARM algorithms (Apriori and FP-Growth) were used to generate rules using the traditional approach of the entire transaction dataset. Next, the selected differential approaches were implemented using both algorithms and results were analysed to compare against the results of the traditional approach. To test for variability in algorithm performance, all experiments were conducted twice. The focus of this chapter is to present the comparison of rule outcomes and algorithm performance between the traditional and differential approaches, based on the results obtained from the second run.

4.1 Traditional Approach

The dataset used contained 252,547 transactions with an average of 3 items per transaction. Five minimum support values (0.001, 0.002, 0.003, 0.004, 0.005) represented as 0.1%, 0.2%, 0.3%, 0.4% and 0.5% respectively were used iteratively as the thresholds to determine the frequent itemsets to consider for generating rules using lift as the metric of interest with a minimum threshold of 1.5. Apriori was found to be faster than FP-Growth in executing, with a total runtime of 298 seconds compared to 319 seconds for FP-Growth. Both algorithms used more time in generating rules at 0.001 minimum support and were faster as the support value increased (see Figure 6). FP-Growth utilized more memory than Apriori across all support thresholds, using an average, using 0.72GB more on average (see Figure 7). Both algorithms generated the same number of rules for all support values, with identical lift scores for each rule. The rules generated using higher minimum support thresholds (0.002, 0.003, 0.004 and 0.005) were all found in the rules generated using the lowest minimum support value (0.001), which had a total of 372 rules. Figure 8 shows a sample of the rules generated.

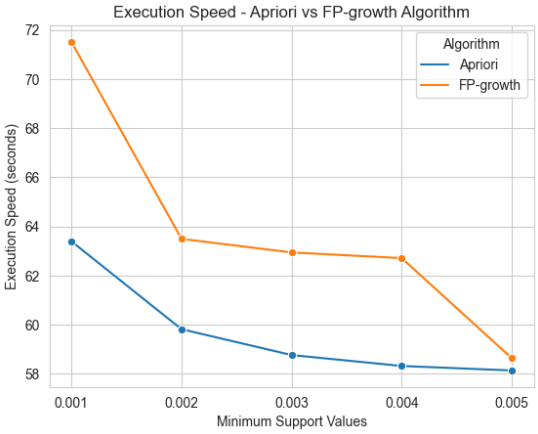


Figure 6: Execution Speed Comparison (Apriori and FP-Growth)

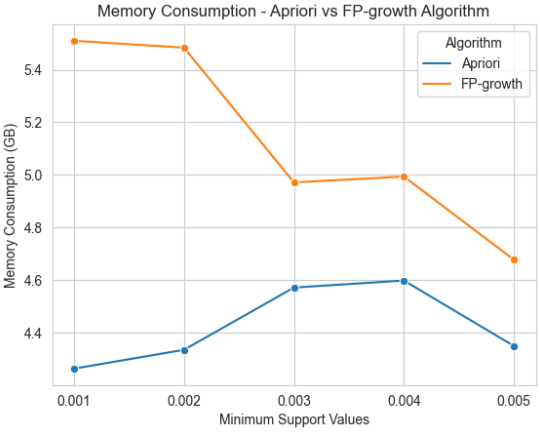


Figure 7: Memory Usage Comparison (Apriori and FP-Growth)

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
79	(MEAT PIE, SAUSAGE ROLL)	(CHICKEN PIE)	0.0045	0.0269	0.0010	0.2231	8.2871	0.0009	1.2525	0.8833
236	(PEAK MILK REFILL POUCH FULL CREAM 400G / 380G)	(MILO SACHET MEDIUM 500G / 550G)	0.0062	0.0054	0.0014	0.2332	42.9227	0.0014	1.2970	0.9828
288	(GOAT MEAT STEW)	(JOLLOF RICE)	0.0064	0.0457	0.0032	0.4935	10.7932	0.0029	1.8840	0.9132
265	(CROISSANT)	(PUFF PUFF)	0.0059	0.1000	0.0012	0.2020	2.0201	0.0006	1.1278	0.5080
19	(JOLLOF RICE)	(CHICKEN STEW)	0.0457	0.0081	0.0040	0.0881	10.8344	0.0037	1.0877	0.9512
329	(FAMILY WHITE BREAD 950G)	(SMALL GREEN APPLES)	0.1447	0.0037	0.0011	0.0077	2.0816	0.0006	1.0040	0.6075
302	(JOLLOF RICE)	(FRIED CHICKEN)	0.0457	0.0191	0.0078	0.1715	8.9863	0.0070	1.1839	0.9313
156	(COKE (PET) 35CL, FANTA ORANGE (PET) 35CL)	(SPRITE (PET) 35CL)	0.0033	0.0070	0.0010	0.3180	45.2220	0.0010	1.4560	0.9811
287	(PUFF PUFF)	(SCOTCH EGG)	0.1000	0.0039	0.0010	0.0102	2.6171	0.0006	1.0064	0.6866
104	(JOLLOF RICE)	(SWAN NATURAL SPRING WATER 75CL)	0.0457	0.0111	0.0012	0.0267	2.3947	0.0007	1.0160	0.6103

Figure 8: Sample of Rules Generated in Traditional Approach

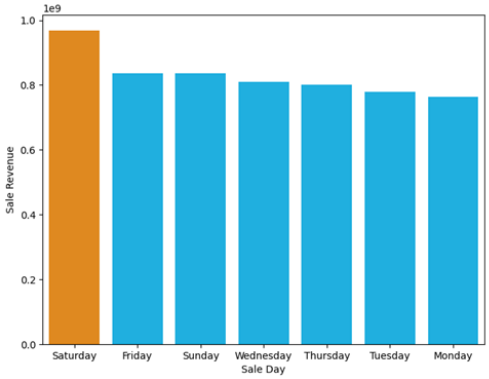
The findings obtained from the traditional approach served as a foundation and informed the implementation strategies used in the experiments involving the differential approaches.

4.2 Differential Approach

The differential factors considered in this experiment were days of the week, months in the year, higher priced items and highest revenue-generating days. The results of these approaches are explained in the following sub-sections. All experiments were run using the same five minimum support threshold values (0.001, 0.002, 0.003, 0.004, 0.005), lift threshold (1.5), and algorithms (Apriori and FP-Growth) used in the traditional approach. Going by the findings from the traditional approach, the results presented in this approach when comparing approaches only state the outcomes using 0.001 as the minimum support threshold. After analysis, FP-Growth performed better at this support threshold (see Appendix F) and is used to compare with the performance of the Apriori algorithm which performed better in the traditional approach. Results in this section focus on the general performance of each differential factor and its comparison with the traditional approach in terms of execution speed, memory consumption and rules generated.

4.2.1 Days of the Week

Experimenting in the order of the particular day in the week that generated the most sale revenue throughout the year (see Figure 9), the dataset was segmented to only include transactions done on each day. Figure 10 shows an overview of the execution speed of the FP-Growth algorithm in generating rules for each of these days across all support thresholds. The algorithm was fastest in generating rules at 0.001 minimum support on Mondays with a runtime of 8.0 seconds, followed by Tuesdays at 8.2 seconds, Wednesdays at 8.4 seconds, Sundays at 8.6 seconds, Fridays at 8.72 seconds, Thursdays at 8.82 seconds and was slowest on Saturdays at 8.9 seconds. A comparison was done to compare the total runtime of all days to the time it took the Apriori algorithm to generate rules at the same support threshold in the traditional approach (see Figure 11). The runtime for all days was found to be better than the traditional approach at 59.9 seconds and 63.3 seconds respectively. In terms of memory used during execution, Saturdays and Sundays used less memory at 4.51GB and 4.65GB respectively, with Thursday using the most at 5.15GB (see Figure 12). When compared to that of the traditional approach, all days combined used an average of 4.9GB compared to 4.2GB of the Apriori algorithm in the traditional approach.



	Sale Day	Sale Revenue
0	Saturday	968,793,362.2692
1	Friday	836,514,544.9493
2	Sunday	836,500,831.5646
3	Wednesday	810,362,871.8220
4	Thursday	799,903,984.0562
5	Tuesday	778,065,885.2708
6	Monday	762,395,354.2048

Figure 9: Sale Revenue by Days

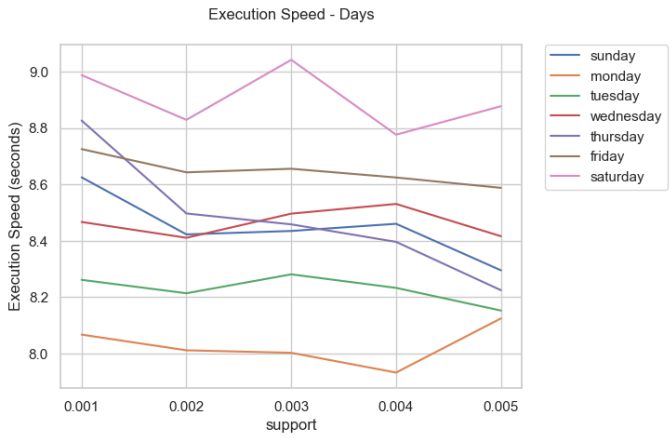


Figure 10: Execution Speed Comparison for All Days

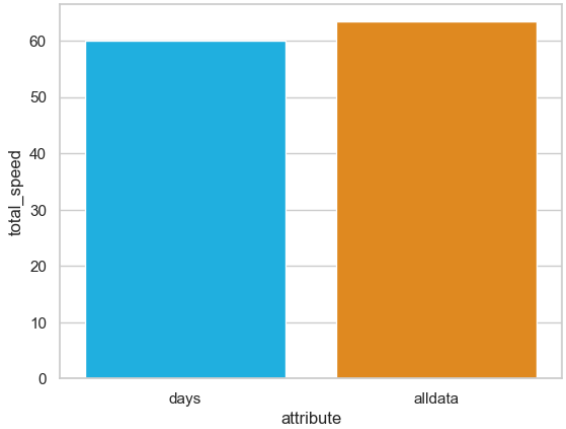


Figure 11: Execution Speed Comparison (Days and Traditional Approach)

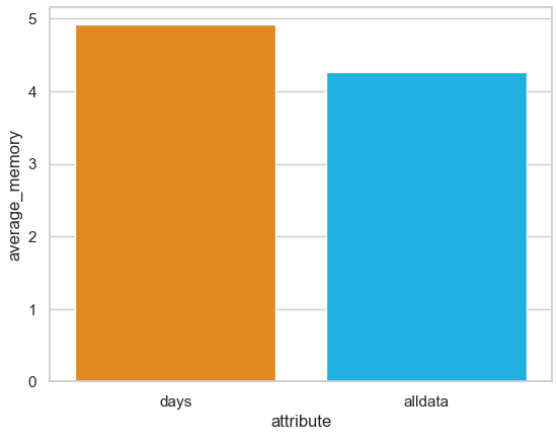


Figure 12: Memory Usage Comparison (Days and Traditional Approach)

An analysis was done to compare the rules generated using the traditional approach with the rules generated using a combination of these days and the lift metric was used to compare the effectiveness of these rules. 218 rules were found in common between both approaches and it was seen that for every rule found in the traditional approach, at least one day had that same rule with a higher lift score than that of the traditional approach. Figure 13 shows a sample of some of these rules and highlights all instances where the rule had a higher lift score.

	Association_Rule	alldata	sunday	monday	tuesday	wednesday	thursday	friday	saturday
0	['COKE ORIGINAL (CAN) 33CL'] => ['FANTA ORANGE CAN 33CL']	71.7845	58.7792	81.3632	74.6441	77.6460	77.3349	66.9357	69.3223
1	['JOLLOF RICE'] => ['WHITE RICE']	2.2922	2.0223	2.1236	2.5146	2.4712	2.0197	2.2649	2.5924
2	['CHICKEN PIE'] => ['PLAIN DOUGHNUT']	5.8660	5.4594	6.6973	4.6002	5.7755	6.2917	6.2754	6.1612
3	['CHICKEN SUYA'] => ['JOLLOF RICE']	9.2565	9.5691	8.7688	8.5413	9.0603	9.1621	9.5921	10.0376
4	['PLAIN DOUGHNUT'] => ['SAUSAGE ROLL']	9.2297	9.9234	10.0223	7.7701	8.5620	8.2638	8.7900	11.6506
5	['HOT DOG'] => ['PLAIN DOUGHNUT']	8.2734	9.0421	9.8303	8.2219	6.8053	7.5730	7.5346	8.9617
6	['SPRITE (PET) 60CL'] => ['FANTA ORANGE (PET) 60CL']	11.0013	10.2349	10.2491	9.1984	11.2816	11.5652	12.5789	11.7853
7	['FANTA ORANGE (PET) 35CL'] => ['COKE (PET) 35CL']	15.8328	15.5457	15.9075	14.2118	15.8752	15.7036	15.9793	16.6372
8	['OFADA SAUCE'] => ['PUFF PUFF']	1.9535	1.9099	1.9077	1.8511	1.8254	2.0460	2.0276	2.1171
9	['COKE (PET) 35CL'] => ['SPRITE (PET) 35CL']	17.4122	17.2506	16.5439	19.9017	17.2774	16.8549	16.9040	16.9185

Figure 13: Similar Rules Analysis (Days and Traditional Approach)

An analysis was done to find instances where rules were found across all days but not in the traditional approach. 308 rules were found to exist in at least one of the days in the week but were not identified in the traditional approach. Figure x shows 10 of these instances. There was no instance where a rule generated (372 total rules) in the traditional approach was not found in any of the days.

	Association_Rule	alldata	sunday	monday	tuesday	wednesday	thursday	friday	saturday
5	['TITUS FISH'] => ['NESTLE H2O (SMALL) 60CL']	NaN	1.9162	NaN	NaN	1.7163	NaN	1.7506	2.1314
6	['JOLLOF RICE', 'CHOPPED FRIED PLANTAIN'] => ['FRIED CHICKEN']	NaN	10.0672	NaN	NaN	NaN	NaN	NaN	NaN
7	['FAMILY WHITE BREAD 950G'] => ['ROSE PLUS WHITE TOILET TISSUE BOULOS - IMPROVED']	NaN	NaN	2.0381	NaN	NaN	NaN	NaN	NaN
8	['FRIED CHICKEN'] => ['JOLLOF RICE', 'CHOPPED FRIED PLANTAIN']	NaN	10.0672	NaN	NaN	NaN	NaN	NaN	NaN
9	['PUFF PUFF', 'SAUSAGE ROLL'] => ['MEAT PIE']	NaN	6.1746	NaN	NaN	NaN	NaN	NaN	NaN
12	['ASUN'] => ['PUFF PUFF', 'JOLLOF RICE']	NaN	13.2686	NaN	NaN	NaN	NaN	NaN	NaN
13	['HOT DOG'] => ['PUFF PUFF', 'MEAT PIE']	NaN	7.1012	NaN	NaN	NaN	NaN	NaN	NaN
14	['CHICKEN SUYA'] => ['MOI-MOI']	NaN	NaN	NaN	NaN	NaN	5.2048	NaN	NaN
15	['CHICKEN ONION SAUCE'] => ['CHOPPED FRIED PLANTAIN']	NaN	8.8952	NaN	NaN	NaN	NaN	NaN	NaN
17	['CHICKEN SUYA', 'JOLLOF RICE'] => ['PUFF PUFF']	NaN	NaN	NaN	NaN	NaN	1.8368	1.9606	2.1408

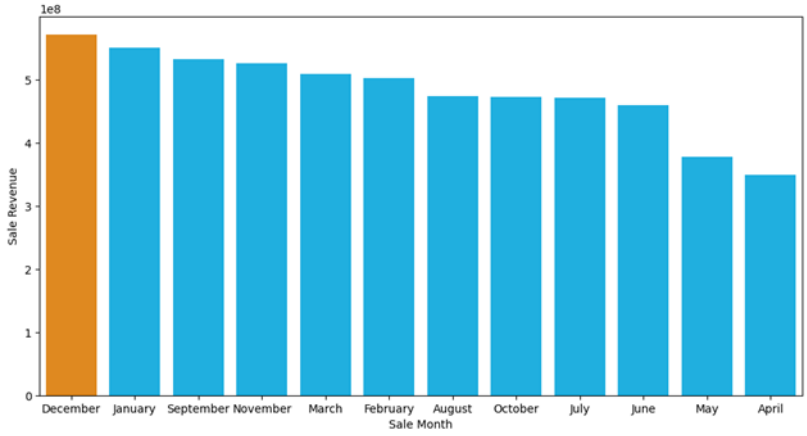
Figure 14: Different Rules Analysis (Days and Traditional Approach)

Table 2: Rules Count (All Days and Traditional Approach)

Min Support	Number of Rules							
	Traditional	Saturday	Sunday	Friday	Wednesday	Thursday	Tuesday	Monday
0.001	372	448	494	366	356	366	362	340
0.002	128	136	150	134	126	124	130	128
0.003	64	76	78	68	60	68	64	66
0.004	36	32	44	36	32	36	36	32
0.005	10	14	18	14	10	14	12	12

4.2.2 Months in the Year

When considering months, going from the highest revenue-generating month to the lowest, (see Figure 15), the dataset was segmented to only include transactions for each of these months and rules were generated accordingly using the FP-Growth Algorithm as most months generated rules faster at 0.001 using this algorithm. Figure 16 shows an overview of the execution speed of the FP-Growth algorithm in generating rules for each month across all minimum support thresholds. The algorithm was fastest in generating rules at 0.001 minimum support from transactions done in April, August, September, and November at 4.2, 4.7, 5.1 and 5.8 seconds respectively. June, October, and March had the highest runtimes of 11.2, 12.5 and 23.1 seconds respectively. A comparison was done to compare the cumulative runtime of all months with the runtime of the Apriori algorithm to generate rules at the same support threshold (0.001) in the traditional approach (see Figure 17). The runtime in the traditional approach was found to be faster, executing at 63.3 seconds compared to 98.3 seconds of the cumulative execution time of all months. In terms of memory used during execution, December used the least memory at 5.19GB, with March using the most at 7.28GB (see Figure 18). When compared to the traditional approach, all months combined used an average of 6.5GB compared to 4.2GB of the Apriori algorithm in the traditional approach.



	Sale Month	Sale Revenue
0	December	571,325,928.3956
1	January	550,197,228.7524
2	September	532,390,672.4072
3	November	525,930,190.8912
4	March	508,932,998.3562
5	February	502,250,564.8934
6	August	473,430,655.5042
7	October	471,658,218.6428
8	July	471,269,179.4179
9	June	458,696,819.1428
10	May	377,697,667.4850
11	April	348,756,710.2482

Figure 15: Sale Revenue by Month

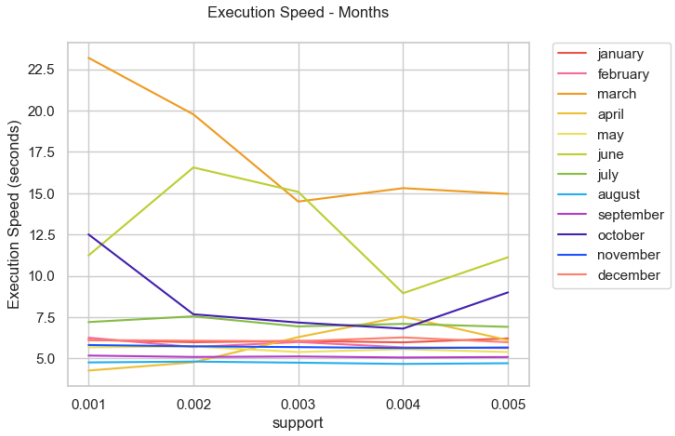


Figure 16: Execution Speed Comparison for All Months

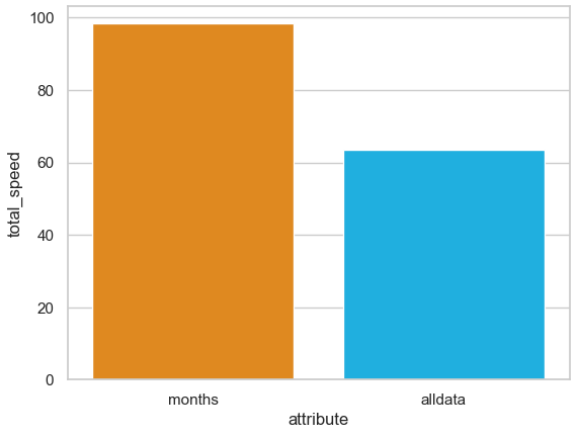


Figure 17: Execution Speed Comparison (Months and Traditional Approach)

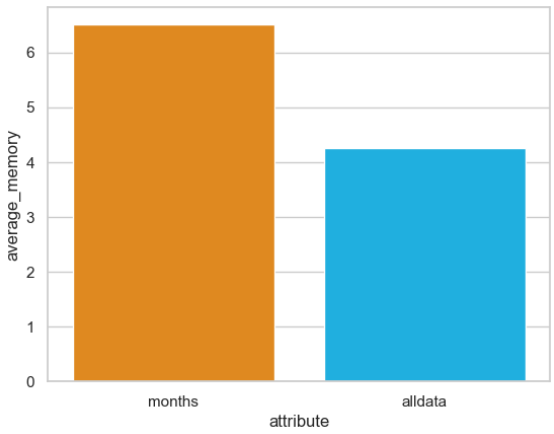


Figure 18: Memory Usage Comparison (Months and Traditional Approach)

An analysis was done to compare the rules generated using the traditional approach with the rules generated using a combination of these months and the lift metric was used to compare the effectiveness of these rules. 28 rules were found in common between both approaches and it was seen that for every rule found in the traditional approach, at least one month had that same rule with a higher lift score than that of the traditional approach. Figure 19 shows a sample of some of these rules and highlights all instances where the rule had a higher lift score.

	Association_Rule	alldata	january	february	march	april	may	june	july	august	september	october	november	december
0	['MEAT PIE'] => ['FISH ROLL/ PIE']	4.8190	5.5121	4.7316	4.7913	4.8383	4.4143	5.2696	3.7595	4.8134	4.7723	4.6108	5.3292	5.0970
1	['CHICKEN PIE'] => ['PLAIN DOUGHNUT']	5.8660	6.4687	5.2896	5.7071	8.8472	5.8687	4.7158	5.8758	5.3946	7.1016	5.0269	5.5456	6.2190
2	['PLAIN DOUGHNUT'] => ['SAUSAGE ROLL']	9.2297	12.3009	7.1656	8.3112	7.1822	9.2268	8.3382	7.7899	11.0841	10.5770	10.5349	9.3026	10.0800
3	['SAUSAGE ROLL'] => ['CHICKEN PIE']	7.3802	6.6502	8.2808	8.3335	6.3804	7.3045	6.0073	6.3715	8.0900	6.7402	8.3807	7.6560	7.7163
4	['FAMILY WHITE BREAD 950G'] => ['THREE CROWN EVAPORATED MILK159ML']	3.3595	4.1413	3.6746	3.7011	1.8413	2.5536	3.2533	2.9089	3.3802	2.8916	3.7032	3.9492	4.2912
5	['SAUSAGE ROLL'] => ['MEAT PIE']	6.2888	6.6593	5.9420	5.8345	7.5357	6.1409	5.6694	4.8799	6.2128	6.3383	6.6524	6.8458	6.2620
6	['PLAIN DOUGHNUT'] => ['CHICKEN PIE']	5.8660	6.4687	5.2896	5.7071	8.8472	5.8687	4.7158	5.8758	5.3946	7.1016	5.0269	5.5456	6.2190
7	['BREAKFAST ROLLS'] => ['FAMILY WHITE BREAD 950G']	2.7485	2.6072	2.7872	3.3677	1.7274	2.2719	2.8188	3.0270	2.9195	3.2623	2.8133	2.6860	2.5282
8	['PLAIN DOUGHNUT'] => ['MEAT PIE']	4.9676	5.0612	4.8743	5.0128	4.9580	4.5974	3.6951	4.0743	5.1221	6.0676	5.7686	5.4279	5.6985
9	['MEAT PIE'] => ['HOT DOG']	5.1638	5.7830	5.1017	4.9882	6.6761	4.5200	4.6093	4.5537	5.7437	4.6298	5.1166	5.1718	5.3456

Figure 19: Similar Rules Analysis (Months and Traditional Approach)

An analysis was done to find instances where rules were found across all months but not in the traditional approach. 1774 rules were found to exist in at least one of the months but were not identified in the traditional approach. Figure 20 shows some of these instances. There was no instance where a rule generated (372 total rules) in the traditional approach was not found among any of the months.

	Association_Rule	alldata	january	february	march	april	may	june	july	august	september	october	november	december
2	['FANTA ORANGE (PET) 35CL'] => ['SPRITE (PET) 35CL', 'COKE (PET) 35CL']	41.1199	NaN	45.9205	NaN	NaN	NaN	NaN	59.0105	NaN	NaN	56.2252	42.3655	32.9351
3	['CHICKEN SUYA'] => ['MEAT PIE']	1.5725	1.7361	1.7188	1.9565	NaN	NaN	NaN	NaN	NaN	NaN	1.9068	NaN	1.7877
9	['MIRINDA PINEAPPLE (PET) 50CL'] => ['PUFF PUFF']	2.4197	NaN	2.3980	2.6113	NaN	NaN	NaN	1.8681	NaN	NaN	1.9526	2.1719	NaN
13	['5 ALIVE PULPY ORANGE (PET) 40CL / 30CL'] => ['PUFF PUFF']	1.5798	NaN	1.5701	NaN	NaN	NaN	NaN	1.6190	2.1566	1.6007	1.7878	1.6399	NaN
18	['FAMILY WHITE BREAD 950G'] => ['MILO SACHET SMALL 200G']	3.1973	NaN	4.6277	NaN	NaN	2.1665	2.9984	NaN	4.2312	NaN	NaN	3.8966	4.2231
32	['GOLDEN PENNY SUGAR CUBES 500g'] => ['FAMILY WHITE BREAD 950G']	2.9416	3.2099	2.9025	3.9452	2.0284	1.7747	3.0014	2.7181	NaN	2.8779	3.2291	3.3304	3.7113
42	['SPRITE (PET) 35CL'] => ['COKE (PET) 35CL', 'FANTA ORANGE (PET) 35CL']	45.2220	NaN	42.3423	NaN	NaN	NaN	NaN	54.2897	NaN	NaN	44.8708	31.9610	35.7631
46	['SPRITE (PET) 35CL', 'FANTA ORANGE (PET) 35CL'] => ['COKE (PET) 35CL']	33.9633	NaN	34.0919	NaN	NaN	NaN	NaN	36.7850	NaN	NaN	40.1650	34.6198	26.8599
53	['TITUS FISH'] => ['FRIED RICE']	10.5141	10.3946	11.2121	8.8980	NaN	37.4366	12.4129	10.5770	8.6840	7.5564	7.8311	9.4893	8.6797
59	['MALTINA CAN 33CL'] => ['AMSTEL CAN MALTA LOW SUGAR 33CL']	14.2624	8.7273	9.6753	16.0242	NaN	19.0882	NaN	NaN	23.7818	28.3345	13.7863	NaN	NaN

Figure 20: Different Rules Analysis (Months and Traditional Approach)

Table 3: Rules Count (All Months and Traditional Approach)

Min Support	Number of Rules												
	Traditional	December	January	September	November	March	February	August	October	July	June	May	April
0.001	372	506	404	568	532	420	454	534	556	542	474	400	792
0.002	128	170	148	188	176	124	140	172	172	160	142	128	174
0.003	64	86	78	98	84	66	82	80	82	86	78	60	104
0.004	36	46	42	50	54	32	40	54	48	58	40	26	60
0.005	10	20	12	24	22	14	16	32	30	28	20	18	38

4.2.3 Higher Priced Items

The dataset for this approach was segmented to only include items in transactions that were priced higher than 75% of the price of items in the whole dataset (4,200.00). Items below this price were removed from transactions before rules were generated using the FP-Growth Algorithm. Figure 21 shows that generating rules from higher-priced items used less time compared to the traditional approach at 45.1 seconds to 63.3 seconds respectively. However, it consumed more memory in the process using 7.2GB compared to 4.2GB used by the Apriori algorithm in the traditional approach (see Figure 22).

Figure 21: Execution Speed Comparison (Expensive and Traditional Approach)

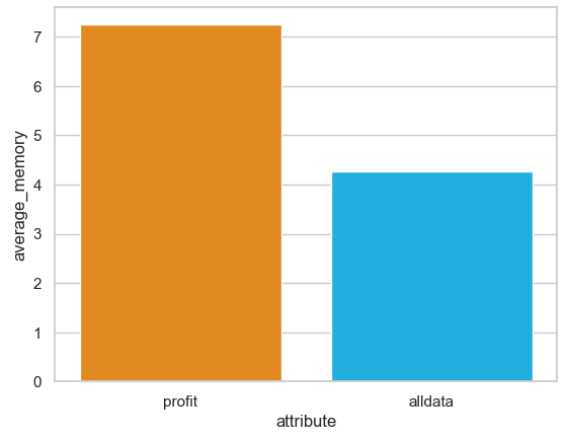


Figure 22: Memory Usage Comparison (Expensive and Traditional Approach)

When comparing the rules generated, 98 rules were found in common between both methods and the rules generated using the traditional approach produced higher lift scores in all 98 instances. Figure 23 shows some of these instances.

	Association_Rule	alldata	profit
0	['FRIED RICE'] => ['FRIED CROAKER FISH']	7.8756	4.1157
1	['CHICKEN SUYA'] => ['JOLLOF RICE']	9.2565	4.8374
2	['JOLLOF RICE'] => ['POTTAGE BEANS']	4.8410	2.5299
3	['FRIED RICE'] => ['GOAT MEAT STEW']	9.9612	5.2056
4	['GOAT MEAT STEW'] => ['FRIED RICE']	9.9612	5.2056
5	['WHITE RICE'] => ['OFADA SAUCE']	18.3782	9.6043
6	['MILO SACHET BIG 1KG'] => ['FAMILY WHITE BREAD 950G']	3.1600	1.6514
7	['FRIED RICE', 'FRIED CHICKEN'] => ['JOLLOF RICE']	6.3474	3.3171
8	['JOLLOF RICE'] => ['OFADA SAUCE']	4.9988	2.6123
9	['FRIED RICE'] => ['TITUS FISH']	10.5141	5.4946
10	['GOAT MEAT STEW'] => ['JOLLOF RICE']	10.7932	5.6405
11	['PEAK MILK REFILL POUCH FULL CREAM 400G / 380G'] => ['MILO SACHET MEDIUM 500G / 550G']	42.9227	22.4310
12	['POTTAGE BEANS'] => ['JOLLOF RICE']	4.8410	2.5299
13	['JOLLOF RICE'] => ['FRIED RICE']	5.3111	2.7755
14	['JOLLOF RICE'] => ['PUFF PUFF', 'FRIED CHICKEN']	7.9316	4.1450
15	['JOLLOF RICE'] => ['FRIED CHICKEN']	8.9863	4.6962
16	['FAMILY WHITE BREAD 950G'] => ['MILO SACHET MEDIUM 500G / 550G']	2.9208	1.5264
17	['FRIED CHICKEN'] => ['FRIED RICE', 'JOLLOF RICE']	13.4285	7.0176
18	['WHITE RICE'] => ['BEEF STEW']	11.3162	5.9137
19	['PUFF PUFF', 'CHICKEN PIE'] => ['MEAT PIE']	5.6293	2.9418

Figure 23: Similar Rules Analysis (Expensive and Traditional Approach)

The same methods used previously were applied to compare the differences between the rules generated by both methods. It was found that there were 174 rules among the higher-priced rules that were not found in the rules generated in the traditional approach (see Figure 24). On the other hand, 274 rules were found in the traditional approach that were not present among the rules generated for the higher-priced items (see Figure 25).

	Association_Rule	alldata	profit
0	['CHOPPED FRIED PLANTAIN'] => ['ASUN', 'JOLLOF RICE']	NaN	7.1318
4	['BEEF STEW'] => ['CHOPPED FRIED PLANTAIN']	NaN	4.6699
7	['JOLLOF RICE', 'CHOPPED FRIED PLANTAIN'] => ['FRIED CHICKEN']	NaN	4.4663
8	['FRIED CHICKEN'] => ['JOLLOF RICE', 'CHOPPED FRIED PLANTAIN']	NaN	4.4663
11	['ASUN'] => ['PUFF PUFF', 'JOLLOF RICE']	NaN	5.6890
12	['ASUN'] => ['VEGETABLE PASTA']	NaN	6.1606
13	['CHICKEN ONION SAUCE'] => ['CHOPPED FRIED PLANTAIN']	NaN	3.1632
18	['CHICKEN ONION SAUCE'] => ['FRIED CHICKEN']	NaN	1.5656
20	['WHITE RICE'] => ['ASUN']	NaN	2.6870
23	['NESTLE H2O (SMALL) 60CL', 'JOLLOF RICE'] => ['TITUS FISH']	NaN	7.1688
26	['FRIED RICE', 'JOLLOF RICE'] => ['TITUS FISH']	NaN	7.0222
29	['CHICKEN SUYA'] => ['FRIED CHICKEN']	NaN	1.8736
31	['CHOPPED FRIED PLANTAIN'] => ['GOAT MEAT STEW']	NaN	4.1470
33	['ASUN', 'JOLLOF RICE'] => ['CHOPPED FRIED PLANTAIN']	NaN	7.1318
35	['JOLLOF RICE'] => ['ASUN', 'CHOPPED FRIED PLANTAIN']	NaN	6.0911
37	['PUFF PUFF', 'JOLLOF RICE'] => ['CHICKEN ONION SAUCE']	NaN	4.7567
38	['TITUS FISH'] => ['POTTAGE BEANS']	NaN	7.2294
41	['JOLLOF RICE'] => ['TITUS FISH', 'FRIED RICE']	NaN	3.5472
42	['FRIED CHICKEN'] => ['FRIED CROAKER FISH']	NaN	2.7204
45	['FRIED CHICKEN'] => ['NESTLE H2O (SMALL) 60CL', 'JOLLOF RICE']	NaN	4.1712

Figure 24: Different Rules Analysis (Expensive > Traditional Approach)

	Association_Rule	alldata	profit
1	['PLAIN DOUGHNUT'] => ['SAUSAGE ROLL']	9.2297	NaN
2	['FANTA ORANGE (PET) 35CL'] => ['SPRITE (PET) 35CL', 'COKE (PET) 35CL']	41.1199	NaN
3	['GOLDEN MORN MAIZE & SOYA PROTEIN 450G'] => ['FAMILY WHITE BREAD 950G']	2.2721	NaN
5	['CHICKEN SUYA'] => ['MEAT PIE']	1.5725	NaN
6	['MIRINDA ORANGE (PET) 50CL'] => ['MEAT PIE']	1.7208	NaN
9	['COKE ORIGINAL (PET) 60CL'] => ['MEAT PIE']	1.9958	NaN
10	['MIRINDA PINEAPPLE (PET) 50CL'] => ['PUFF PUFF']	2.4197	NaN
14	['NESCAFE BREAKFAST 3IN1 32G'] => ['FAMILY WHITE BREAD 950G']	2.6296	NaN
15	['FISH ROLL/ PIE'] => ['SAUSAGE ROLL']	12.2603	NaN
16	['NESTLE H2O (SMALL) 60CL'] => ['FRIED RICE']	1.5296	NaN
17	['5 ALIVE PULPY ORANGE (PET) 40CL / 30CL'] => ['PUFF PUFF']	1.5798	NaN
19	['PEAK MILK POWDER SACHET 16G'] => ['MILO 20g (ENERGY FOOD DRINK)']	43.4242	NaN
21	['JOLLOF RICE'] => ['FRIED CHICKEN']	8.9863	4.6962
22	['MOI-MOI'] => ['PUFF PUFF']	2.4914	NaN
24	['FAMILY WHITE BREAD 950G'] => ['MILO SACHET SMALL 200G']	3.1973	NaN
25	['SAUSAGE ROLL'] => ['FISH ROLL/ PIE']	12.2603	NaN
27	['FRIED RICE'] => ['NESTLE H2O (SMALL) 60CL']	1.5296	NaN
28	['CROISSANT'] => ['RAISINS']	40.2760	NaN
30	['FRUIT BREAD 550g'] => ['FAMILY WHITE BREAD 950G']	1.8493	NaN
32	['GOLDEN PENNY SUGAR CUBES 500g'] => ['FAMILY WHITE BREAD 950G']	2.9416	NaN

Figure 25: Different Rules Analysis (Traditional Approach > Expensive)

Table 4: Rules Count (Expensive and Traditional Approach)

Min Support	Number of Rules	
	Traditional	Expensive
0.001	372	272
0.002	128	96
0.003	64	54
0.004	36	34
0.005	10	30

4.2.4 Specific Day Rules

To compare the differences between rules and performance using transactions on a specific day, the dataset was filtered to generate rules from transactions done on 24th December 2020, as this was the highest selling day (see Figure 26). The difference between execution times was huge as rules generated for this day took just 0.6 seconds compared to the 63.3 seconds in the traditional approach (see Figure 27). However, this used significantly more memory at 6.8GB compared to the 4.2GB used when the entire dataset was used in the traditional approach (see Figure 28).

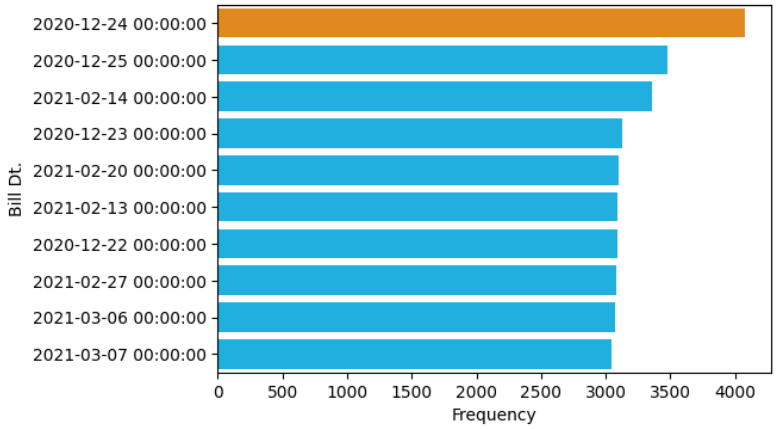


Figure 26: Sale Revenue by Day

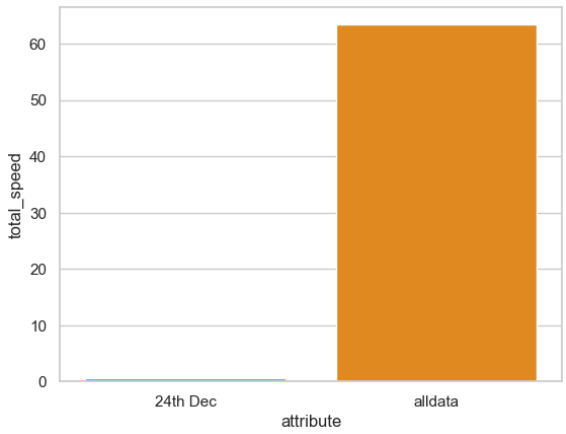


Figure 27: Execution Speed Comparison (24th Dec and Traditional Approach)

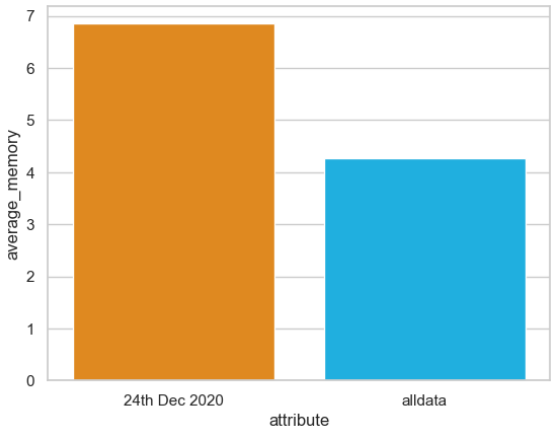


Figure 28: Memory Usage Comparison (24th Dec and Traditional Approach)

In terms of rules generated, there were 166 rules in common with varying lift scores. Among these rules, 136 of them had higher lift scores on the 24th of December, in contrast to just 30 having higher lift scores from the traditional approach. Figure 29 highlights some of these instances.

	Association_Rule	alldata	24th Dec
0	['COKE ORIGINAL (CAN) 33CL'] => ['FANTA ORANGE CAN 33CL']	71.7845	56.4103
1	['CHICKEN SUYA'] => ['JOLLOF RICE']	9.2565	6.7901
2	['PLAIN DOUGHNUT'] => ['SAUSAGE ROLL']	9.2297	22.9167
3	['HOT DOG'] => ['PLAIN DOUGHNUT']	8.2734	21.1538
4	['SPRITE (PET) 60CL'] => ['FANTA ORANGE (PET) 60CL']	11.0013	8.4615
5	['CHOPPED FRIED PLANTAIN'] => ['PUFF PUFF']	1.8606	1.6858
6	['JOLLOF RICE'] => ['POTTAGE BEANS']	4.8410	10.1852
7	['MIRINDA ORANGE (PET) 50CL'] => ['MEAT PIE']	1.7208	2.2665
8	['SCOTCH EGG'] => ['PUFF PUFF']	2.6171	5.7471
9	['PUFF PUFF'] => ['5 ALIVE PULPY ORANGE (PET) 40CL / 30CL']	1.5798	1.7241
10	['BREAKFAST ROLLS'] => ['FAMILY WHITE BREAD 950G']	2.7485	6.2678
11	['BOILED YAM'] => ['EGG SAUCE']	118.6882	220.0000
12	['COKE ORIGINAL (PET) 60CL'] => ['MEAT PIE']	1.9958	2.3384
13	['MILO SACHET BIG 1KG'] => ['FAMILY WHITE BREAD 950G']	3.1600	3.1339
14	['MEAT PIE'] => ['HOT DOG']	5.1638	6.0440
15	['EGG SAUCE'] => ['BOILED YAM']	118.6882	220.0000
16	['JOLLOF RICE'] => ['OFADA SAUCE']	4.9988	12.2222
17	['SPRITE (PET) 60CL'] => ['MEAT PIE']	2.1102	3.0220
18	['MEAT PIE'] => ['PUFF PUFF']	1.7016	2.4836
19	['FRIED RICE'] => ['TITUS FISH']	10.5141	20.3704

Figure 29: Similar Rules Comparison (24th Dec and Traditional Approach)

Table 5: Rules Count (24th Dec and Traditional Approach)

Min Support	Number of Rules	
	Traditional	24th December 2020
0.001	372	4420
0.002	128	694
0.003	64	278
0.004	36	144
0.005	10	94

This chapter concisely stated the results of the experiments conducted in the modeling phase of this dissertation without any speculative interpretations. Notes on the types of experiments were given as well as the reasons why they were conducted. This next chapter will provide a detailed interpretation of these results including what they mean, their importance and their relevance to the research problem.

5 DISCUSSION

The findings from this research have provided insight into the effects and performance differences between the traditional and differential approaches of generating association rules from offline retail transactions. However, elements of these results should be interpreted with caution due to the unique nuances of the dataset and limitations of the current research. This chapter reflects on the research process and summarizes key findings from the results of the experiments conducted, the implications of their interpretation in practical applications, as well as limitations that may have influenced the results.

Generally, there was variability in the execution process of both algorithms in terms of runtime and memory usage after running tests multiple times. This can be attributed to the order in which the transactions are processed as it was observed that each time the dataset was loaded in the working environment, the transactions were always in a different order. This variability in performance can also be attributed to other internal software processes making use of system resources during executions. Nonetheless, FP-Growth performed better in most instances when the minimum support threshold was set to 0.001. This is in line with the consensus in the literature that FP-Growth is generally faster than Apriori due to the advanced data structure it uses to store candidate itemsets.

The results from the experiments conducted indicate that certain differential approaches in mining association rules from retail transactions can be incorporated to improve the quality and significance of rules generated and also optimise computational resources used.

When transactions were segmented into days of the week, the results showed that the same rules generated when using the whole dataset were all also found in at least one day segment with a higher lift score, there were also a significant number of niche rules that were missed by the traditional approach. This suggests that mining association rules using days as a differential factor increases the quality and significance of rules and provides specific context into when that rule can be exploited to potentially yield better results. The results of this approach also showed that the total time it took to generate rules for all seven days was shorter than when the entire dataset was used in the traditional approach.

The results when transactions were segmented by months showed a similar outcome in terms of rule significance as every rule generated in the traditional approach was also found in at least one month with a higher lift score. However, execution speed and memory used were worse in comparison to the traditional approach and this may indicate that the nature of the dataset when segmented using this factor impacts the rule-generating process of the algorithm.

Using higher-priced items to generate rules showed that although it executed faster, similar rules generated were not as significant because there were rules generated in the traditional approach that were not found using this factor. However, some rules found using this factor were also absent

from the traditional approach. This suggests that focusing on just expensive items may provide similar results in less time but the differences between both approaches are marginal.

Lastly, the results when rules were generated from a single day (24th December 2020) in comparison to the whole dataset met expectations in terms of execution speed due to the huge disparity in the size of the datasets. However, this surprisingly, used more system memory and may be a result of other system processes running concurrently and may also be because it was the last experiment done in a sequence of other memory-consuming experiments. Furthermore, a lot of rules generated were a result of those rules happening once among transactions on this day (confidence of 1) and may not be applicable as this can be attributed to pure chance. This approach however can be combined with domain knowledge to identify rules that may have resulted from seasonal spending habits.

Overall, the findings from this study provide new insights into the relationship between several differential factors in generating item association rules from retail transactions. While previous research has focused on mining rules using higher support thresholds on large transactions to optimise for performance, these results demonstrate that segmenting the dataset using differential factors with even lower support thresholds can yield similar and more significant results. These results can thereby be considered in real-life retail applications for optimising marketing strategies and physical store layouts. However, it is important to note that the generalizability of these results is limited by the customer habits specific to the transactions in this dataset and may not be representative of every ARM implementation from retail transactions.

6 CRITICAL EVALUATION

This chapter evaluates to what extent the research objectives and research questions stated in earlier chapters have been achieved and answered.

6.1 Objectives

Objective 1: Collect offline retail transactions datasets from retail business and pre-process them for the generation of association rules.

- This objective was fully achieved in the data understanding and data preparation phases of the methodology (section 3.2 and 3.3). Datasets were collected and explored to understand their content and distribution frequencies. Furthermore, the dataset was cleaned to handle errors and validity issues. New attributes were generated from existing ones for experiments in the modelling phase.

Objective 2: To develop an effective association rule mining system to extract association rules from retail transactions data using traditional and differential approaches.

- This objective was fully achieved and results from implementations and experiments done using the developed rule mining system can be seen in section 4.

Objective 3: Conduct experiments to evaluate the performance impact of differential approaches in mining association rules in comparison to the traditional approach.

- This objective was fully achieved and evaluation results and findings from the experiments conducted using both approaches were presented and discussed in detail in section 4 and section 5.

6.2 Research Questions

- **Research Question 1:** How can association rule mining techniques be effectively applied to retail transactions data to uncover interesting item patterns and associations?

  This study showed that ARM algorithms (Apriori and FP-Growth) from the mlxtend Python library can be used to efficiently mine association rules from retail transactions data with user defined support thresholds and other evaluation metrics (section 3 and 4)

- **Research Question 2:** How can differential rule mining approaches such as time hierarchies, item value and specific day sales improve the significance of association rules?

  The experiments conducted in this study and results presented in section 4 and 5, show how day and month time hierarchies improve the quality and significance of rules generated compared to the rules generated using the traditional approach.

- **Research Question 3:** How can differential rule mining approaches improve the performance of association rule mining algorithms in terms of speed and memory usage?

  The experiments conducted also show that differential approaches have the potential to optimize computational resources used by ARM algorithms in generating rules. However, factors such as randomness and other system processes mean that this is not a conclusive result (section 5)

7 CONCLUSION

This research aimed to develop an efficient association rule mining system and investigate the extent to which differential approaches can improve the overall rule extraction process in terms of rule interestingness and performance from an offline retail transaction dataset.

The dataset was from a retail business in Nigeria that contained 252,547 transactions with 10,217 items over one year. An ARM system with customizable parameters was developed to identify rules using traditional and differential approaches. It can be concluded that segmenting the dataset by days and months are important factors to consider in generating association rules from retail transactions. The results using these factors showed a positive effect on the quality and significance of rules generated using the FP-Growth algorithm. The rules generated using these factors showed the segment of transactions where these rules were dominant and could be exploited for better results in implementations in the retail industry. The same rules generated using the traditional approach were found using these approaches but with the added advantage of the context of when these rules are stronger. However, performance evaluation in terms of execution speed and memory usage could not be concluded due to several combined factors that lead to variability in performance. Differential approaches focused on mining rules using only expensive items and on specific days are not as reliable due to their poor representation of the entire customer base but may be investigated further by domain experts to understand the significance of rules.

This research has shown that incorporating days and months as separate factors into the rule-mining process of retail transactions can improve the quality and significance of rules. This has the potential to improve retail business operations and customer satisfaction through better inventory management and optimized marketing strategies. Given the lack of extant research on the application of differential approaches in mining association rules, future work can be done to test for generalizability due to the limitations of this study, while using this research as a baseline. Furthermore, further research on the different outcomes of this experiment on multi-store retail store businesses can be done to evaluate the impact of these approaches on different customer segments and this can be incorporated into the ARM system developed in this study.

Word count (main body of the report): 10266

REFERENCES

Aggarwal, C. C., Bhuiyan, M. A. and Hasan, M. Al, 2014. Frequent Pattern Mining Algorithms: A Survey. *In*: *Frequent Pattern Mining*. Cham: Springer International Publishing, 19–64.

Agrawal, R., Imieliński, T. and Swami, A., 1993. Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, 22 (2), 207–216.

Akram, U., Fülöp, M. T., Tiron-Tudor, A., Topor, D. I. and Căpușneanu, S., 2021. Impact of Digitalization on Customers' Well-Being in the Pandemic Period: Challenges and Opportunities for the Retail Industry. *International Journal of Environmental Research and Public Health*, 18 (14), 7533.

Alfiqra and Khasanah, A. U., 2020. Implementation of Market Basket Analysis based on Overall Variability of Association Rule (OCVR) on Product Marketing Strategy. *IOP Conference Series: Materials Science and Engineering*, 722 (1), 012068.

Ali, M., 2023. *Association Rule Mining in Python Tutorial | DataCamp* [online]. Datacamp. Available from: https://www.datacamp.com/tutorial/association-rule-mining-python [Accessed 1 Mar 2023].

Bagui, S. and Dhar, P. C., 2019. Positive and negative association rule mining in Hadoop's MapReduce environment. *Journal of Big Data* [online], 6 (1), 75. Available from: https://doi.org/10.1186/s40537-019-0238-8.

Chang, H. H., Wong, K. H. and Fang, P. W., 2014. The effects of customer relationship management relational information processes on customer-based performance. *Decision Support Systems*, 66, 146–159.

Deng, Z.-H. and Lv, S.-L., 2014. Fast mining frequent itemsets using Nodesets. *Expert Systems with Applications*, 41 (10), 4505–4512.

Du, J., Zhang, X., Zhang, H. and Chen, L., 2016. Research and improvement of Apriori algorithm. *In*: *2016 Sixth International Conference on Information Science and Technology (ICIST)*. IEEE, 117–121.

Duggirala, R. and Narayana, P., 2013. Mining Positive and Negative Association Rules Using CoherentApproach.

E. Bala Krishna A. Nagaraju, B. R., 2015. A Survey on Effective Mining of Negative Association Rules from Huge Databases. *International Journal of Computer Sciences and Engineering* [online], 3 (9), 220–223. Available from: https://www.ijcseonline.org/full_paper_view.php?paper_id=674.

Fournier-Viger, P., Lin, J. C.-W., Kiran, R. U., Koh, Y. S. and Thomas, R., 2017. A survey of sequential pattern mining. *Data Science and Pattern Recognition*, 1 (1), 54–77.

Giampaolo Rodola, 2023. *psutil documentation — psutil 5.9.5 documentation* [online]. Available from: https://psutil.readthedocs.io/en/latest/ [Accessed 26 May 2023].

Hameli, K., 2018. A literature review of retailing sector and business retailing types. *ILIRIA International Review*, 8 (1), 67–87.

Han, J., Pei, J. and Tong, H., 2022. *Data mining: concepts and techniques*. Morgan kaufmann.

Han, J., Pei, J. and Yin, Y., 2000. Mining frequent patterns without candidate generation. *ACM SIGMOD Record*, 29 (2), 1–12.

Hidber, C., 1999. Online association rule mining. *In*: *Proceedings of the 1999 ACM SIGMOD international conference on Management of data*. New York, NY, USA: ACM, 145–156.

Kavitha, B., Konda, S., Reddy, B. and Govardhan, D., 2011. Mining Negative Association Rules. *International Journal of Engineering and Technology*, 3, 100–105.

Kavitha, M. and Selvi, Ms. S. T. T., 2016. Comparative Study on Apriori Algorithm and Fp Growth Algorithm with Pros and Cons. *In*: .

Koh, Y. S. and Ravana, S. D., 2016. Unsupervised Rare Pattern Mining: A Survey. *ACM Trans. Knowl. Discov. Data* [online], 10 (4). Available from: https://doi.org/10.1145/2898359.

Lisa, I., Stefano, N., Chris, M., Zineb, S. and Omar Frikha, 2022. *The Future of Traditional Retail in Africa* [online]. Boston Consulting Group. Available from: https://www.bcg.com/publications/2022/the-future-of-traditional-retail-in-africa [Accessed 26 May 2023].

Luna, J. M., Fournier-Viger, P. and Ventura, S., 2019. Frequent itemset mining: A 25 years review. *WIREs Data Mining and Knowledge Discovery*, 9 (6).

M. Shridhar, M. P., 2017. Survey on Association Rule Mining and Its Approaches. *International Journal of Computer Sciences and Engineering* [online], 5 (3), 129–135. Available from: https://www.ijcseonline.org/full_paper_view.php?paper_id=1223.

Mahmood, S., Shahbaz, M. and Guergachi, A., 2014. Negative and Positive Association Rules Mining from Text Using Frequent and Infrequent Itemsets. *The Scientific World Journal*, 2014, 1–11.

Nurzani, Z. and Tania, K. D., 2020. Analysis of Transactions 212 Mart Kuto Palembang to Find Frequent Patterns Among Itemset Using Association Rule Mining. *In*: *Proceedings of the Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019)*. Paris, France: Atlantis Press.

Papavasileiou, V. and Tsadiras, A., 2011. Time Variations of Association Rules in Market Basket Analysis. *In*: Iliadis, L., Maglogiannis, I., and Papadopoulos, H., eds. *12th Engineering Applications of Neural Networks (EANN 2011) and 7th Artificial Intelligence Applications and Innovations (AIAI)* [online]. Corfu, Greece: Springer, 36–44. Available from: https://hal.inria.fr/hal-01571480.

Raja, V., 2017. *CRISP-DM still a leader in data mining models* [online]. New Zealand: Stellar. Available from: https://stellarconsulting.co.nz/articles/crisp-dm-still-a-leader/ [Accessed 26 May 2023].

Raschka, S., 2022. *association_rules: Association rules generation from frequent itemsets* [online]. mlxtend: Wisconsin. Available from:

http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/association_rules/ [Accessed 1 Mar 2023].

Savasere, A., Omiecinski, E. and Navathe, S., 1999. Mining for strong negative associations in a large database of customer transactions. *In*: *Proceedings 14th International Conference on Data Engineering*. IEEE Comput. Soc, 494–502.

Schröer, C., Kruse, F. and Gómez, J. M., 2021. A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, 181, 526–534.

Solanki, S. K. and Patel, J. T., 2015. A Survey on Association Rule Mining. *In*: *2015 Fifth International Conference on Advanced Computing & Communication Technologies*. IEEE, 212–216.

Telikani, A., Gandomi, A. H. and Shahbahrami, A., 2020. A survey of evolutionary computation for association rule mining. *Information Sciences*, 524, 318–352.

Tugba Sabanoglu, 2022. *Projected retail sales growth worldwide from 2020 to 2025* [online]. USA: Statista. Available from: https://www.statista.com/statistics/232347/forecast-of-global-retail-sales-growth/ [Accessed 26 May 2023].

Wu, X., Zhang, C. and Zhang, S., 2004. Efficient mining of both positive and negative association rules. *ACM Transactions on Information Systems*, 22 (3), 381–405.

Yan, X., Zhang, C. and Zhang, S., 2009. CONFIDENCE METRICS FOR ASSOCIATION RULE MINING. *Applied Artificial Intelligence*, 23 (8), 713–737.

Yudhistyra, W. I., Risal, E. M., Raungratanaamporn, I. and Ratanavaraha, V., 2020. Using Big Data Analytics for Decision Making: Analyzing Customer Behavior using Association Rule Mining in a Gold, Silver, and Precious Metal Trading Company in Indonesia. *International Journal of Data Science*, 1 (2), 57–71.

APPENDIX A – PROJECT PROPOSAL

Bournemouth University

Department of Computing and Informatics

2022-23 Academic Year Individual Masters Project

Project Proposal Form

Please refer to the **Project Handbook Section 4** when completing this form. Note that your proposal should be your own original work and you must cite sources in line with university guidance on **referencing and plagiarism**[1].

Degree Title: **MSc Data Science and Artificial Intelligence**	**Student's Name:** David Anda
	Supervisor's Name: Marcin Budka
	Project Title/Area: A Context-Based Association Rule Mining Recommendation Engine From Offline Retail Transactions

Section 1: Project Overview

1.1 Problem definition - use one sentence to summarise the problem:

Retail companies are constantly looking for ways to better understand customer behaviour and purchase patterns to optimise product offerings, increase revenue, improve store layout/product placement, and efficiently manage inventory and this research will delve into non-traditional approaches of association rule mining to understand if they create better product associations.

1.2 Project description - briefly explain your project:

The purpose of this project is to develop an association rule mining recommendation engine for a Nigerian retail business using context-based approaches like seasonality, business-specific peak and off-peak periods, pricing groups, product categories, and profitability of products. These approaches would look to generate faster and more efficient association rules to create precise marketing strategies to maximise profits.

[1] https://libguides.bournemouth.ac.uk/study-skills-referencing-plagiarism

Bournemouth University

Department of Computing and Informatics

2022-23 Academic Year Individual Masters Project

1.3 Background - please provide brief background information, e.g., client, problem domain, and make reference to the literature (minimum 4-5 sources):

Market basket or association analysis is an unsupervised data mining technique used in retail industries to understand buying habits of customers and ultimately looks to answer questions about what products are usually bought together from previous transactions (Han et al. 2012; Kotu and Deshpande 2019). This information gives insights to the stakeholders and decision-makers about their customers and helps them craft selective marketing strategies and better optimize the placement of products in their physical stores – that would in turn improve customer relationships and improve sales. The concept of market basket analysis goes far beyond the supermarket scenario from which it takes its name and can be applied in other industries as it is simply the analysis of a collection of items to find exploitable relationships (Loshin 2013). An example of market basket analysis in action is if most of the time, customers that buy computers, often buy antivirus software, putting these two products close to each other would help increase the sale of both products . A limitation of current techniques of association rule mining is algorithm performance, execution time and computational cost when working with extremely large datasets (Hossain et al. 2019). There is also the issue of generating too many rules that are less relevant and difficult to explain when using the frequency-based approach of purchased products in the whole dataset as the base point for generating these rules (Chun-sheng and Yan 2014). Hence, my research would focus on these limitations and look to explore other base points that scale down the dataset using context-based approaches in the creation of association rules from real-life retail transaction data.

1.4 Aims and objectives – what are the aims and objectives of your project? should be specific and measurable:

This project aims to investigate to what extent specific context-based approaches can optimise and improve product association rule mining in terms of execution time, usefulness of rules and algorithm performance to improve marketing strategies and revenue generation opportunities in a chain store fast-moving consumer goods (FMCG) retail business.

1.5 Research Questions

Can using context-based approaches like seasonality, product profitability and specific retail key performance indicators generate more efficient product association rules than a product purchase frequency-based approach?

Bournemouth University

Department of Computing and Informatics

2022-23 Academic Year Individual Masters Project

Section 2: Artefact

2.1 What is the artefact that you intend to produce?

A well-documented set of scripts using Python (Jupyter Notebook) that would allow for a comparative performance analysis of context-based and frequency-based approaches in the generation of product association rules from offline retail transactions data.

2.2 How is your artefact actionable (i.e., routes to implementation and exploitation in the technology domain)?

This artefact would be actionable to the customer relationship management and marketing teams of retail companies with access to POS transactional data. It would enable them plug in periodic data and discover patterns and recommendations on relationships between products and possible cross-selling and marketing strategies and sale techniques to generate more revenue.

Section 3: Evaluation

3.1 How are you going to evaluate your project artefact?

This artefact would be evaluated using the quality/interestingness and speed of model execution in the generation of product association rules using context-based and frequency-based approaches and through client usage as I would be testing this solution with the retail store in Nigeria that is providing the dataset for this project. The success of potential marketing campaigns and increase in revenue using these product recommendations would determine the quality of this research and allow me identify improvement opportunities.

3.2 How does this project relate to your MSc Programme and your degree title outcomes?

This project relates to my programmes as I would be applying the knowledge gained from my coursework in processing and analysing large databases and using machine learning to discover underlying patterns, evaluate its performance using statistical metrics, provide recommendations and ultimately use computing to solve a business problem at scale.

3.3 What are the risks in this project and how are you going to manage them?

There is minimal risk involved with this project as the only obstacle would be the inability to get the data from the retail company in Nigeria in time for analysis and development. I am currently in contact with the

Bournemouth University

Department of Computing and Informatics

2022-23 Academic Year Individual Masters Project

IT department and on-track to extract the data by the end of the year after end of year sales have been recorded.

Section 4: References

4.1 Please provide references if you have used any.

Chun-sheng, Z. and Yan, L., 2014. Extension of local association rules mining algorithm based on apriori algorithm. *2014 IEEE 5th International Conference on Software Engineering and Service Science*, 340-343.

Concepts and Methods. *In:* Han, J., Kamber, M. and Pei, J., eds. 2012/01/01/. *Data Mining (Third Edition)* [online]. Boston: Morgan Kaufmann, 243-278.

Hossain, M., Sattar, A. S. and Paul, M. K., 2019. Market basket analysis using apriori and FP growth algorithm, *2019 22nd International Conference on Computer and Information Technology (ICCIT)* (pp. 1-6): IEEE.

Kotu, V. and Deshpande, B., 2019. Chapter 6 - Association Analysis. *In:* Kotu, V. and Deshpande, B., eds. 2019/01/01/. *Data Science (Second Edition)* [online]. Morgan Kaufmann, 199-220.

Loshin, D., 2013. Chapter 17 - Knowledge Discovery and Data Mining for Predictive Analytics. *In:* Loshin, D., ed. 2013/01/01/. *Business Intelligence (Second Edition)* [online]. Morgan Kaufmann, 271-286.

Section 5: Academic Practice and Ethics

Please delete as appropriate.

5.1 Have you made yourself familiar with, and understand, the University guidance on referencing and plagiarism? **Yes**

5.2 Do you acknowledge that this project proposal is your own work and that it does not contravene any academic offence as specified in the University's regulations? **Yes**

Note: Please complete the research ethics checklist once the proposal has been approved by your supervisor.

Bournemouth University

Department of Computing and Informatics

2022-23 Academic Year Individual Masters Project

Section 6: Proposed Plan (please attach your Gantt chart below)

APPENDIX B – BOURNEMOUTH UNIVERSITY RESEARCH ETHICS CHECKLIST

Bournemouth University

Research Ethics Checklist

About Your Checklist	
Ethics ID	46525
Date Created	23/11/2022 18:54:00
Status	Approved
Date Approved	24/11/2022 10:40:38
Date Submitted	23/11/2022 18:59:40
Risk	Low

Researcher Details	
Name	David Anda
Faculty	Faculty of Science & Technology
Status	Postgraduate Taught (Masters, MA, MSc, MBA, LLM)
Course	MSc Data Science & Artificial Intelligence

Project Details	
Title	A Context-Based Association Rule Mining Recommendation Engine From Offline Retail Transactions
Start Date of Project	01/02/2023
End Date of Project	19/05/2023
Proposed Start Date of Data Collection	01/01/2023
Supervisor	Marcin Budka
Approver	Marcin Budka

Summary - no more than 600 words (including detail on background methodology, sample, outcomes, etc.)

Retail companies are constantly looking for ways to better understand customer behavior and purchase patterns to optimize product offerings, increase revenue, improve store layout/product placement, and efficiently manage inventory and this project will investigate and perform an analysis of different approaches to creating product pairs (association rule s) to understand if they create better product associations and do it in less time and using less computation power.

None of the filter questions apply to my study

I am confirming that my proposed project does not:

- Involve human participants
- Involve the use of human tissue

Page 1 of 2 Printed On 07/03/2023 12:37:34

- Involve medical research requiring NHS ethical / REC Approval
- Involve the use of animals (or tissues/fluids derived from animals)
- Involve access to identifiable personal data for living individuals not already in the public domain
- Involve increased danger of physical or psychological harm for researcher(s) or subject(s)
- Raise any ethical issues associated with the use of genetically modified organisms

On this basis, my proposed project does not require a formal ethics review.

If any changes to the project involve any of the criteria above, I undertake to resubmit the project for formal ethical approval.

APPENDIX C – FIRST PROGRESS REVIEW REPORT

Department of Computing and Informatics

Postgraduate Project First Progress Review

To be completed and signed by the Supervisor and Student during week **commencing 6 March 2023**.

Student: David Anda	**Supervisor: Marcin Budka**

Assessment

1. Definition of the problem *Has the problem, research aims, and questions been defined, has the artefact been identified and have objectives been set?*	Yes
Comments:	
2. Review of literature and related work *Is there evidence of appropriate research?*	Yes
Comments:	
3. Methodology and Artefact *Is there evidence of appropriate analysis of the problem and design of a solution and appropriate evaluation?*	To some extent
Comments:	
4. Dissertation *Have sections of the dissertation been written and has the Supervisor seen these?*	Yes
Comments:	
5. Planning & Progress *Is there an acceptable plan for this project and is it being followed?*	To some extent
Comments: Slightly behind but will catch up.	
6. Proposal & Online Ethics Checklist *Are proposal and ethic checklist submitted? Are they approved?*	Yes, both are submitted and approved
7. Overall Assessment	Requires Minor Improvement
Signed: Supervisor: …Marcin Budka…….. Student: ……David Anda..... Date: ……07/03/2023……	

- Supervisor to retain the signed form and supply the student with a copy if required.
- Supervisor to upload the form on Brightspace and grade as *Satisfactory, Requires Major Improvement, Requires Minor Improvement, Unsatisfactory or Invalid*.
- Supervisor to notify the Project Coordinator if the student is at risk of failing the project or not engaging.

APPENDIX D – LIST OF CONTENT OF LARGE FILES

After Unzipping the file. The following content would be found:

- artefact.ipynb (file) : The main script to reproduce results of the experiments conducted in this study.
- requirements.txt (file): contains the dependencies and libraries required to reproduce the experiments conducted in this study.
- project_files (folder): contains the datasets used in this study.

Guide:

All experiments were conducted on a Jupyter Notebook on a Windows 11 PC. You will only need to change the working directory of the dataset to reproduce these results.