Market Basket Anaysis

Mette

2024-02-25

Table of contents

Figure 1: Market basket analysis

1 Introduction

In today's fast-paced and ever-evolving retail landscape, where consumers have a plethora of options at their fingertips, gaining a deep understanding of customer behavior has become paramount for the sustained success of any business. In this fiercely competitive environment, companies must constantly adapt and innovate to stay ahead of the curve and meet the evolving needs and preferences of their target audience.

One powerful tool in a retailer's arsenal for deciphering customer preferences and purchasing patterns is the analysis of transaction data. By meticulously examining the transactions made by customers, retailers can unearth valuable insights that provide a comprehensive understanding of consumer behavior, preferences, and trends. These insights, in turn, serve as the foundation upon which informed business decisions can be made, ultimately leading to enhanced customer satisfaction, increased profitability, and sustainable growth.

In this project, we embark on a journey of exploration and discovery as we delve into a rich dataset containing grocery store transactions. Our objective is clear: to glean actionable insights that will not only shed light on the intricacies of consumer behavior but also serve as the guiding beacon for strategic decision-making within the retail domain. Through rigorous analysis and interpretation of transaction data, we seek to unlock hidden patterns, identify emerging trends, and uncover correlations that hold the key to unlocking the full potential of the grocery store business.

2 Analysis

The primary objective of this project is to conduct exploratory analysis of grocery store transactions to identify patterns, trends, and associations among purchased items. By analyzing transaction data, we seek to gain insights into customer behavior, preferences, and purchasing habits. These insights can then be used to inform various business decisions, such as targeted marketing campaigns, product recommendations, and inventory management strategies.

2.1 CRISP-DM metodology

The Cross-Industry Standard Process for Data Mining (CRISP-DM) offers a systematic framework for conducting data mining projects. It encompasses six distinct stages: comprehending the business context, understanding the data, preparing the data, developing models, evaluating these models, and implementing them.

2.1.1 Business understanding

The goal of this analysis is to identify trends based on collected data about customer transactions. This can lead to important insights that can be used for marketing efforts, product placement and layout optimization, product recommendations, and inventory management. Overall, market basket analysis can provide businesses with valuable insights into customer preferences and behavior, which they can use to make informed decisions and improve their business results.

2.1.2 Data Understanding

Upon initial inspection, we have performed a thorough check for missing values in the dataset, and we can confirm that there are none present. This ensures that the dataset is complete and ready for further analysis. Furhter more to understand the trend we will investigate the top 10 most purchased items and their frequency.

2.1.2.1 Reading libraries

```
#|output: false
#|warnings: false
#|codefode: true


pacman::p_load("tidyverse", "magrittr", "nycflights13", "gapminder",
               "Lahman", "maps", "lubridate", "pryr", "hms", "hexbin",
               "feather", "htmlwidgets", "broom", "pander", "modelr",
               "XML", "httr", "jsonlite", "lubridate", "arules",
               "arulesViz", "datasets", "gclus", "DT", "arules")
```

2.1.2.2 Importing data

The dataset is provided and loaded in basketformat.

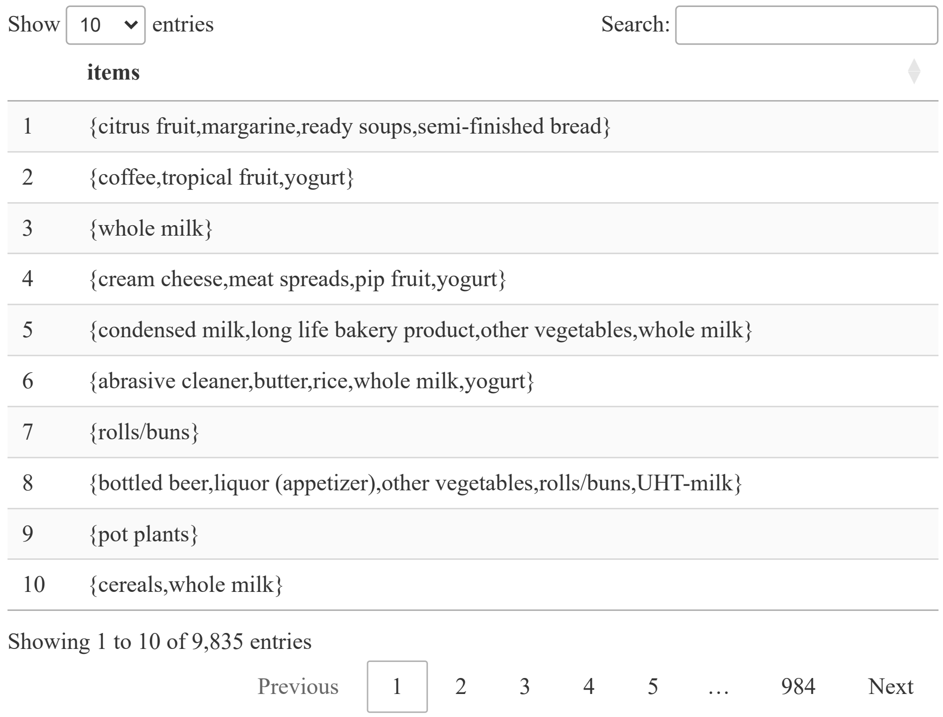
```
# format = basket ---------------------------------------------------------

g <- read.transactions("groceries.csv", format = "basket", sep = ",", cols = NULL)
```

This creates a dataset with 9.835 observations that we can inspect. This specific dataset has allready been cleaned and is ready for analysis.

```
#vi skaber en interaktiv tabel

datatable(as(g, "data.frame"))
```



	items
1	{citrus fruit,margarine,ready soups,semi-finished bread}
2	{coffee,tropical fruit,yogurt}
3	{whole milk}
4	{cream cheese,meat spreads,pip fruit,yogurt}
5	{condensed milk,long life bakery product,other vegetables,whole milk}
6	{abrasive cleaner,butter,rice,whole milk,yogurt}
7	{rolls/buns}
8	{bottled beer,liquor (appetizer),other vegetables,rolls/buns,UHT-milk}
9	{pot plants}
10	{cereals,whole milk}

Showing 1 to 10 of 9,835 entries

Upon initial inspection, we have performed a thorough check for missing values in the dataset, and we can confirm that there are none present. This ensures that the dataset is complete and ready for further analysis. Furhter more to understand the trend we will investigate the top 10 most purchased items and their frequency.

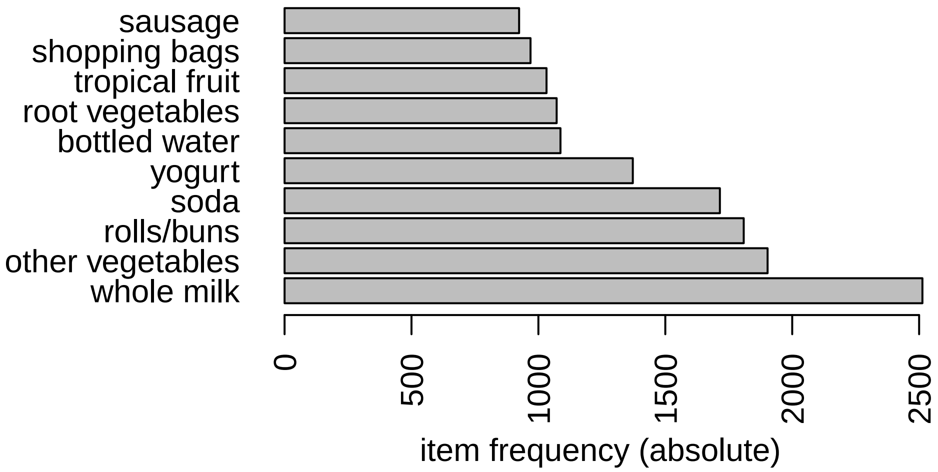
```
# Tæl antallet af manglende værdier i objektet g
missing_values <- sum(is.na(g@data))

# Udskriv antallet af manglende værdier
cat("Missing values in the dataset g:", missing_values, "\n")
```

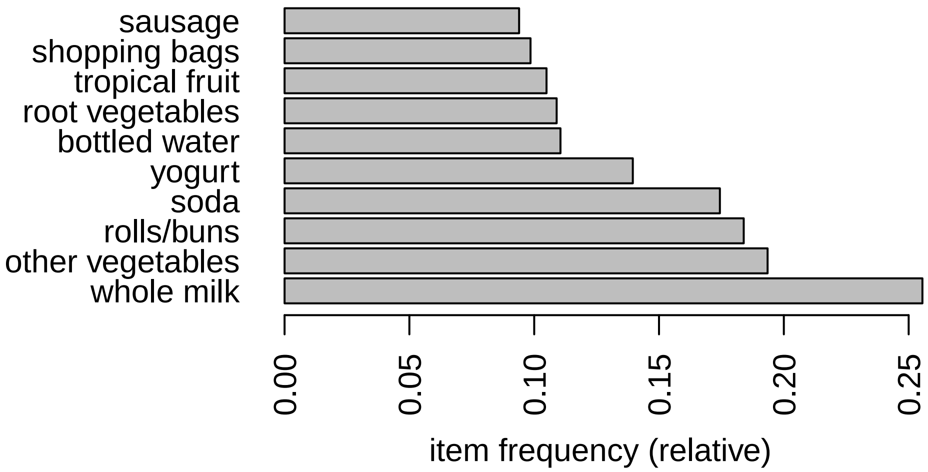
```
Missing values in the dataset g: 0
```

```
# item frequency -----------------------------------------------------------

itemFrequencyPlot(g, topN=10, type="absolute", horiz=TRUE)
```



```
itemFrequencyPlot(g, topN=10, type="relative", horiz=TRUE)
```



2.1.3 Data Preparation

After importing the data, it has been expected and is ready for analysis

2.1.4 Modelling

2.1.4.1 Exploratory analysis

Exploratory analysis of transaction data involves examining patterns and relationships between items frequently purchased together. The purpose is to discover interesting and significant associations that can provide insights into consumer behavior and preferences.

One of the most common approaches to exploratory analysis of transaction data is market basket analysis. Using techniques like the Apriori algorithm, we can identify frequent combinations of items, also known as associations or rules. These rules specify which items are often bought together and can be crucial for understanding customer purchasing behavior.

The Apriori algorithm works by iteratively generating candidate itemsets of increasing size based on the frequent itemsets found in the previous iteration. It prunes the search space by using the “apriori” property, which states that if an itemset is infrequent, all its supersets will also be infrequent. This property helps reduce the number of candidate itemsets that need to be examined.

During exploratory analysis, we can evaluate several aspects of the derived rules, including:

1. **Support**: This measures the frequency of a given combination of items relative to the total number of transactions in the dataset. High support indicates that the items are frequently purchased together.
2. **Confidence**: This measures the probability that a particular item is also purchased when another item is already in the basket. High confidence suggests a strong association between the items.
3. **Lift**: Lift indicates how much more likely it is for the items to be purchased together compared to what would be expected if their purchases were independent of each other. Lift values above 1 indicate a positive association between the items.

By analyzing these metrics, we can identify important associations between items that can inform strategies for product placement, cross-selling, marketing, and other business decisions.

In the following sections, we will conduct a detailed exploratory analysis of our transaction data to discover interesting associations between items and draw meaningful conclusions about consumer preferences and behavior.

Here we are focusing on the top 5 rules.

```
# Oprette et eksempel på transaktionsdata fra datasættet g
rules <- apriori(g, parameter = list(supp = 0.001, conf = 0.8, minlen = 2,
                                     maxlen = 4), control = list(verbose=FALSE)
                 )
rules
# Vis de udledte regler
summary(rules)
```

```
#for nemmere visning af de første 5 regler

for (i in 1:5) {
  antecedents <- labels(lhs(rules[i]))
  consequents <- labels(rhs(rules[i]))
  support <- quality(rules[i])$support
  confidence <- quality(rules[i])$confidence
  lift <- quality(rules[i])$lift

  cat("Rule", i, ":\n")
  cat("Antecedents (If): ", paste(antecedents, collapse = ", "), "\n")
  cat("Consequents (Then): ", paste(consequents, collapse = ", "), "\n")
  cat("Support: ", support, "\n")
  cat("Confidence: ", confidence, "\n")
  cat("Lift: ", lift, "\n\n")
}
```

```
Rule 1 :
Antecedents (If):  {liquor,red/blush wine}
Consequents (Then):  {bottled beer}
Support:  0.001931876
Confidence:  0.9047619
Lift:  11.23527

Rule 2 :
Antecedents (If):  {cereals,curd}
Consequents (Then):  {whole milk}
Support:  0.001016777
Confidence:  0.9090909
Lift:  3.557863

Rule 3 :
Antecedents (If):  {cereals,yogurt}
Consequents (Then):  {whole milk}
Support:  0.001728521
Confidence:  0.8095238
Lift:  3.168192

Rule 4 :
Antecedents (If):  {butter,jam}
Consequents (Then):  {whole milk}
Support:  0.001016777
Confidence:  0.8333333
Lift:  3.261374

Rule 5 :
Antecedents (If):  {bottled beer,soups}
Consequents (Then):  {whole milk}
Support:  0.001118454
Confidence:  0.9166667
Lift:  3.587512
```

2.1.5 Evaluation

In this section, we evaluate the performance of the association rules generated by the Apriori algorithm. The evaluation is based on several key metrics, including support, confidence, and lift. These metrics help us understand the significance and reliability of the discovered associations between antecedents (If) and consequents (Then).

Support

Support measures the frequency of occurrence of a particular itemset in the dataset. For example, in Rule 1, the support value of 0.00193 indicates that approximately 0.193% of all transactions contain both “liquor” and “red/blush wine” in the antecedents.

Confidence

Confidence measures the strength of the association between the antecedents and consequents. For instance, in Rule 2, the confidence value of 0.909 suggests that when “cereals” and “curd” are purchased together (antecedents), there’s a 90.9% chance that “whole milk” (consequent) will also be purchased.

Lift

Lift quantifies the degree of dependency between the antecedents and consequents while considering the baseline probability of purchasing the consequents. A lift value greater than 1 indicates that the antecedents and consequents are positively correlated. For instance, in Rule 3, the lift value of 3.168 suggests that the likelihood of purchasing “whole milk” increases by approximately 3.168 times when both “cereals” and “yogurt” are bought together, compared to when “whole milk” is purchased independently.

Overall, the association rules exhibit high support, confidence, and lift values, indicating strong and significant associations between the antecedents and consequents. These findings can be valuable for retail businesses in understanding customer purchasing behavior and optimizing product placement and marketing strategies.

2.1.6 Implementation

Based on the rules discovered in the modelling section, the following implementation suggestions can be considered:

1. **Product Placement Optimization**: Use the association rules to strategically place related products near each other in physical stores or online platforms. For example, if Rule 1 suggests that “liquor” and “red/blush wine” are frequently purchased together with “bottled beer,” consider placing them in close proximity on shelves or suggesting them as related items in online product listings.

2. **Targeted Marketing Campaigns**: Leverage the association rules to tailor marketing campaigns and promotions based on the identified associations between products. For instance, if Rule 2 indicates a strong association between “cereals” and “curd” leading to the purchase of “whole milk,” you could create targeted promotions or discounts for customers purchasing these items together.

3. **Cross-Selling and Upselling**: Train sales representatives or develop algorithms to recommend additional products to customers based on their current selections. For example, if a customer adds “butter” and “jam” to their cart (as per Rule 4), you could suggest adding “whole milk” as well.
4. **Inventory Management**: Optimize inventory stocking levels based on the discovered associations. Ensure that frequently co-purchased items are adequately stocked to meet customer demand and minimize stockouts. Conversely, consider bundling or promoting items that are less frequently purchased together to increase their sales.
5. **Customer Segmentation**: Use association rules to segment customers based on their purchasing behavior and preferences. Develop targeted marketing strategies or loyalty programs tailored to each segment to enhance customer satisfaction and retention.
6. **Dynamic Pricing**: Adjust pricing strategies for associated products to maximize revenue and profitability. For example, consider offering discounts or promotions on complementary items to incentivize larger purchases.
7. **Product Bundling**: Bundle related products together and offer them at a discounted price to encourage customers to purchase multiple items simultaneously. Ensure that the bundled products align with the associations identified in the association rules.

3 Conclusion

In conclusion, our analysis of grocery store transactions using the Apriori algorithm has provided valuable insights into consumer behavior and purchasing patterns. By uncovering significant associations between items, we have gained a deeper understanding of customer preferences and tendencies within the retail landscape. The high support, confidence, and lift values observed in the association rules signify strong and meaningful relationships between various products, offering actionable opportunities for retailers to optimize their business strategies.

The findings from our analysis have several implications for retailers seeking to enhance customer satisfaction, increase profitability, and drive sustainable growth. From strategic product placement and targeted marketing campaigns to optimized inventory management and dynamic pricing strategies, there are numerous avenues through which retailers can leverage these insights to stay ahead of the competition and meet the evolving needs of their customers.

Moving forward, continuous monitoring and analysis of transaction data will be essential for retailers to adapt to changing consumer preferences and market dynamics. By embracing data-driven decision-making processes and harnessing the power of advanced analytics techniques, retailers can unlock new opportunities for innovation and differentiation in an increasingly competitive marketplace.

In essence, our exploration of grocery store transactions has not only provided valuable insights into consumer behavior but has also underscored the importance of leveraging data and analytics to drive strategic decision-making in the retail industry. By staying attuned to customer needs and preferences, retailers can position themselves for long-term success in today's dynamic retail landscape.