**MARKET BASKET ANALYSIS**

**SIMULATED COLES DATA**

**Table of Contents**

[Executive Summary 1](#_Toc32181049)

[Introduction 1](#_Toc32181050)

[Description of the dataset 2](#_Toc32181051)

[Original Data 2](#_Toc32181052)

[Appendix 3](#_Toc32181053)

[Appendix I – Transactional variables 3](#_Toc32181054)

[Appendix II –Demographical and Socio-Economical Features 3](#_Toc32181055)

[Appendix II – Basket Items 4](#_Toc32181056)

# **Executive Summary**

This report provides an analysis of the Simulated Coles data which consists of 58,100 observations and 53 variables. The aim of the analysis is to explore questions about which products customers are buying and what kind of customers are shopping at Coles. The answers to these questions can be used to enhance customer experience and/or to maximize profits which can be achieved by improving store layouts, offering discounts or marketing different products together, and determining the best place to shelf the specific products.

After cleaning and preprocessing the data, Clustering analysis and Market Basket analysis have been applied to the data. Clustering analysis results revealed that there 4 groups of customers based on income, age and spending value. These 4 groups are named **Elderly High Spenders, High Spending Youth, Young & Broke and Rich & Wise Youth. These groups are named based on their characteristics of the customers in the groups. Market Basket Analysis results uncovered that there is a high correlation between products like bread, milk & banana and nappies & baby food. Some less obvious product combinations were fish, vegetables & household cleaners and coffee, bread, vegetables, frozen meals & household cleaners.**

**This report also offers a detailed explanation of data cleaning and analytical processes, data mining techniques, analysis outcome, and suggestions for future analysis.**

# **Introduction**

Coles is one of the leading supermarkets in Australia, generally dealing with the provision of everyday goods. The emergence of competitors in the market has meant that the competition has become much fiercer and It is crucial for Coles to take strategic actions to beat the competition. One way to do that is to use the data Coles has been collecting from its customers and analyze data to gain a competitive edge. This report aims to use unsupervised knowledge discovery to find hidden patterns in customers’ transactions and to form customer segments. We’ll be seeking a better understanding of two aspects:

* **Customer Segments**: - What are the different segments or groups and their demographic as well as socio-economic features.
* **Products Patterns**: - Popularity of the products, what kind of products are being bought together, the likelihood of two or more products being purchased together.

By better understanding these two aspects and finding answers to the questions related to these two aspects, Coles would be able to form sophisticated strategies that would help Coles to lead the competition. A better understanding of the customer segments would help Coles create more effective and successful market campaigns and design eye-catchy catalogs of the products whereas better knowledge of product patterns would allow coles to improve in-store layouts and product placements.

# **Description of the dataset**

## Original Data

The original dataset used in this analysis was provided into spreadsheet(.xlsx) format. The spreadsheet consists of:

* 58,100 observations(rows)
* 53 variables(columns)
  + 3 Transcational Variables. *[Appendix I]*
  + 6 Variables representing Demographics and Socio-Economical features of customers. *[Appendix II]*
  + 44 Basket Items. *[Appendix III]*

The table below offers a brief description of each variable(For detailed table go to Appendix)

|  |  |  |
| --- | --- | --- |
| **Variable** | **Type** | **Description** |
| ***ReceiptID*** | Ordinal | Transaction identification number |
| ***Value*** | Continuous | Dollar amount of the transaction |
| ***pmethod*** | Nominal | Payment method: 1 = Cash, 2 = Card, 3 = Eftpos, 4 = Other |
| ***sex*** | Binary | 1 = male, 2 = female |
| ***homeown*** | Nominal | If they own a house: 1 = Yes, 2 = No, 3 = Unknown |
| ***income*** | Continuous | Customer’s income per annum in dollars |
| ***age*** | Discrete | Age last birthday |
| ***PostCode*** | Nominal | Post code of customer’s current address |
| ***nchildren*** | Ordinal/Discrete | Number of children of the customer |
| **Basket Items** | Binary | Supermarket items: 0 = not purchased, 1 = purchased |

***Table 1 Variable Description of Original Data***

# **Appendix**

**[Red cell color suggests that data is dirty or missing in that column]**

**[Numbers in Quality of Data for one variable is independent from other variables]**

## Appendix I – Transactional variables

|  |  |  |  |
| --- | --- | --- | --- |
| **Transactional Data** | | | |
| **Variable** | **Type** | **Description** | **Quality of Data** |
| ReceiptID | Numeric - Unique Key | Unique transaction ID | 9 duplicate values |
| Value | Numeric - Continuous | Value of the transaction | No missing values, outliers detected |
| pmethod | Numeric - Categorical | Payment Method  (1 = Cash, 2 = Credit card, 3 = EftPOS, 4 = Other) | 97 erroneous entries [≈0.17%] |

## Appendix II –Demographical and Socio-Economical Features

|  |  |  |  |
| --- | --- | --- | --- |
| **Demographical & Socio-Economical Data** | | | |
| **Variable** | **Type** | **Description** | **Quality of Data** |
| sex | Numeric - Categorical- Binary | Customer’s Gender  (1 = Male, 2 = Female) | No missing values or outliers |
| homeown | Numeric - Categorical | House Ownership  (1 = Yes, 2 = No, 3 = Unknown) | 99 erroneous entries [≈0.17%] |
| income | Numeric - Continuous | Customer’s Income in dollars (Per Annum) | 1 missing value [≈ 0.0017%], Outliers detected |
| age | Numeric - Continuous | Customer’s Age | 1 missing value [≈ 0.0017%], Outliers detected |
| PostCode | String - Categorical | Customer’s Postal Code | 9858 missing values + Erroneous entries [≈17%] |
| nchildren | Numeric - Discrete | No of Children that a customer has | 2 missing value + Erroneous entries detected [≈0.19%] |

## Appendix II – Basket Items

|  |  |  |
| --- | --- | --- |
| **Basket Items** | | |
| **Variable** | **Type** | **Quality of Data** |
| fruit | String | 10 erroneous entries [≈0.017%] |
| freshmeat | Binary | No missing values or outliers |
| dairy | Binary | No missing values or outliers |
| MozerallaCheese | Binary | No missing values or outliers |
| cannedveg | Binary | 1 missing value [≈ 0.0017%] |
| cereal | Binary | 9 missing values [≈0.015%] |
| frozenmeal | Binary | No missing values or outliers |
| frozendessert | Binary | No missing values or outliers |
| pizzabase | Binary | 1 missing value [≈ 0.0017%] |
| TomatoSauce | Binary | No missing values or outliers |
| frozen fish | Binary | No missing values or outliers |
| bread | Binary | No missing values or outliers |
| milk | Binary | 1 missing value [≈0.0017%] |
| softdrink | Binary | No missing values or outliers |
| fruitjuice | Binary | 10 erroneous entries [≈0.017%] |
| confectionary | Binary | 1 missing value [≈ 0.0017%] |
| fish | Binary | No missing values or outliers |
| vegetable | Binary | No missing values or outliers |
| energydrink | Binary | No missing values or outliers |
| tea | Binary | No missing values or outliers |
| coffee | Binary | No missing values or outliers |
| laundrypowder | Binary | No missing values or outliers |
| householcleaners | Binary | No missing values or outliers |
| corn chips | Binary | No missing values or outliers |
| Frozen yogurt | Binary | No missing values or outliers |
| Chocolate | Binary | No missing values or outliers |
| Olive Oil | Binary | No missing values or outliers |
| Baby Food | Binary | No missing values or outliers |
| Napies | Binary | No missing values or outliers |
| banana | Binary | No missing values or outliers |
| cat food | Binary | No missing values or outliers |
| dog food | Binary | No missing values or outliers |
| mince | Binary | No missing values or outliers |
| Sunflower Oil | Binary | No missing values or outliers |
| chicken | Binary | No missing values or outliers |
| vitamins | Binary | No missing values or outliers |
| deodorants | Binary | No missing values or outliers |
| dishwashingliquid | Binary | No missing values or outliers |
| onions | Binary | No missing values or outliers |
| lettuce | Binary | No missing values or outliers |
| KitKat | Binary | No missing values or outliers |
| TeaTowel | Binary | No missing values or outliers |
| Scones | Binary | No missing values or outliers |