**Data Science Postgraduate Project - COSC2667**

**For Coles Super Market**

**Specification of data science project**

**scope and deliverables**

**Name:** Shamini Puthooppallil Baby

**Student ID:** S3674381

# PROJECT DESCRIPTION AND OBJECITVES

The purpose of this project is to build a customer segmentation profile to derive insights. In this project, customer segmentation analysis will be performed on data relating to Coles Impulse product customers [1]. These customers will be segment based on common characteristics.

# BACKGROUND

Customer segmentation is the practice of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing which relies on identifying key differentiators that divide customers into groups that can be targeted. Information such as a customers' demographics (i.e. gender, age, religion, ethnicity, income, education level, etc.), geographic (where they live and work), psychographic ((i.e. social class, lifestyle and personality characteristics, etc.) and behavioural (i.e. interest, spending habit, consumption, usage and desired benefits, etc.) tendencies are considered when determining customer segmentation practices [2].

Coles is an Australian supermarket, retail and consumer services chain which are accumulating, millions of sales transactions, and customer information in their day-to-day business, which are stored in the databases. These databases are collected with valuable information and can be directly applied for making an intelligent business decision which would certainly benefit from customer segmentation. Coles currently has in place the Flybuys rewards program (Australia’s most popular loyalty program) which, can be a customer profile that basically allows for each customer’s transactional information to be pursued [3].

The main goal of retailers is to provide best customer services by knowing their needs and preferences and it is essential for a retailer to predict and find out most profitable customers who account for the major portion of their future profits. Retailers recognized that valuable, non-trivial useful information can be extracted from large retail data, which helps decision maker to take vital decision for business operations. Customer purchasing behaviour is considered significant, based on past sales transactions. Customer segmentation is one of the fundamental tasks, which has been more widely used to identifying right customers, knowing their needs and offering right services at the right time are the main goals of retailers which lead a retail organization to employ and plan a clear strategy for treating different types of customers.

# 

# PROJECT TIMELINE

# 

# METHODOLOGY

The main objectives of the projects are break down into the following categories:

* **Data extraction** – understand data and identify the right fields

This project utilised Coles’ data over the period of the previous 12 months, which ranged from Jan 2018 to Jan 2019. The data used throughout this project came from three basic tables. The first stored information relating to Coles’ products, the second has information relating to Coles’ customer transactions, and the third contains information relating to Coles’ customer-specific attributes (such as demographics, etc.). SQL and R software were used to query and analyse the data, respectively.

* **Research** - determine an appropriate segmentation technique to apply

**K-means**

The K-means clustering algorithm is used to find groups which have not been explicitly labelled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets. Once the algorithm has been run and the groups are defined, any new data can be easily assigned to the correct group [4].

**Hierarchical clustering**

Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom.

* **Data pre-processing** – formatting the data to required format

Prior to analysis, the data was filtered based on number of criteria.

1. Christmas and Easter week were excluded
2. Low sales items were excluded
3. Total sale amount, a filter was applied to remove high value and low value customers.
4. The products which are not sold more than 200 Coles stores excluded.
5. There are 18 sub categories under impulse products and the seasonal products will be removed from the analysis as this will be considered as outlier for the model.

To accomplish the aim of this project, the following factors were created from the previously mentioned basic tables:

1. Impulse product category sales (amount in $)
2. Total Impulse & overall sales amount in $
3. Total discount applied over Impulse products.
4. Average Impulse & overall sales amount (per spend) in $
5. Proportion of $ amount spent on Impulse over overall Coles purchases
6. Number of visits to the Impulse and Coles overall
7. Customer age, gender, affluence, CFSS and life stage

* **Clustering** – apply segmentation techniques
* **Profiling** – identify meaningful segments within impulse

# Deliverables

1. The key insights from past impulse sales
2. The customer segmentation model and profile
3. Model validation
4. Principal component analysis
5. Document describing the steps adopted in each and every step.

# REFERENCES

1. <https://en.wikipedia.org/wiki/Impulse_purchase>
2. <shopify.com.au/encyclopedia/customer-segmentation>
3. <https://en.wikipedia.org/wiki/Flybuys_(Australia)>
4. <https://www.r-bloggers.com/k-means-clustering-in-r/>