RETAIL SUPERMARKET: CUSTOMER ANALYSIS

CAPSTONE PROJECT REPORT

भारतीय प्रबंध संस्थान कोषिक्कोड

Indian Institute of Management Kozhikode

Globalizing Indian Thought

PROFESSIONAL CERTIFICATE PROGRAMME IN ADVANCED ANALYTICS AND BUSINESS INTELLGENCE BATCH 1 (GROUP 14)

GROUP 14 MEMBERS

Group Details	Name	Email ID
14	Vishnuraj Rajendran	vishnurajrajendran17@gmail.com
	Vignesh Aravind	vignesharavind1993@gmail.com
	Vineeta Ann George	vineetaanngeorge@gmail.com
	Vinith Vijayapalan Paramel	vinithvijay@gmail.com
	Yatika Gupta	yatikagupta24@gmail.com
	Sukhbir Singh	sukhu1872@yahoo.com

PROJECT GUIDE

Prof. Sreejesh S
Programme Chair and Associate Professor, Marketing Management

TABLE OF CONTENTS

RETAIL SUPERMARKET: CUSTOMER ANALYTICS

Introduction

Background

The retailer was founded in late 1970s and operated across North American continent. The company primarily sells products from categories including wine, meat, fish, fruit, confectionary and gold. The retailer was a pioneer in using data to inform its product stocking levels using ledgers keeping track of store transactions, product placement and customer surveys. This helped the retailer to increase revenue with an efficient cost structure enabling their market expansion across the region at a rapid pace.

Problem Statement

Based on the success of data analytics shown by other competitors, the retailer is looking to include analytics into their day-to-day business decisions.

The primary objective being to capture insights and trends relating to:

- Customer behaviour mainly spending habits
- Product preferences
- Promotional responses
- Additional factors influencing future growth

The end goals include improved shopping experience for customer, undertake decisions at par with competitors, capture opportunities to boost revenue. Etc.

Significance of the Study

This study aims at generating insights and action areas for the retailer so as to achieve an understanding of the customer needs and aspirations. The detailed descriptive analytics will help in capturing underlying insights which acts as a guide in realigning the retailer's current strategies and improve the customer enhancement and experience campaigns. Additionally, the customer prediction and segmentation will help in generating a baseline of the customer expectations which can guide the retailer to optimize their current processes and workflows.

Methodology

The entire analysis activity will be performed using the below set of tools and techniques.

- **Tableau** – Exploratory Data Analysis
- **R language and associated packages** – Customer spend predication and segmentation
- **Prediction algorithms** – Linear regression, Decision Tree, Random Forest, XGBoost

Data Summary

The data under consideration has the following characteristics:

- 2240 data rows across 29 columns
- Data consists of following column categories
    - Customer biographical information
        - Year_Birth
        - Education
        - Marital_Status
        - Income
        - Kidhome
        - Teenhome
        - Dt_Customer
        - Recency
    - Customer Shopping information
        - MntWines
        - MntFruits
        - MntMeatProducts
        - MntFishProducts
        - MntSweetProducts
        - MntGoldProds
        - NumDealsPurchases
        - NumWebPurchases
        - NumCatalogPurchases
        - NumStorePurchases
        - NumWebVisitsMonth
        - AcceptedCmp3
        - AcceptedCmp4
        - AcceptedCmp5
        - AcceptedCmp1
        - AcceptedCmp2
        - Complain
- The 3 columns (Z_CostContact, Z_Revenue, Response) which have been excluded.
- **Special handling:**
    - Income which are blanks or are in an extreme range have been added as “Exceptions” in the Income Bins.
    - Income for which data is available if binned as below for EDA purposes:

[Income] > 1000 AND [Income] < 10000	1K - 10K
[Income] >= 10000 AND [Income] < 25000	10K - 25K
[Income] >= 25000 AND [Income] < 50000	25K - 50K
[Income] >= 50000 AND [Income] < 75000	50K - 75K
[Income] >= 75000 AND [Income] < 90000	75K - 90K
[Income] >= 90000 AND [Income] < 110000	90K - 110K
[Income] >= 110000 AND [Income] < 125000	110K - 125K
[Income] >= 125000 AND [Income] < 150000	125K - 150K
[Income] >= 150000 AND [Income] < 200000	150K - 200K

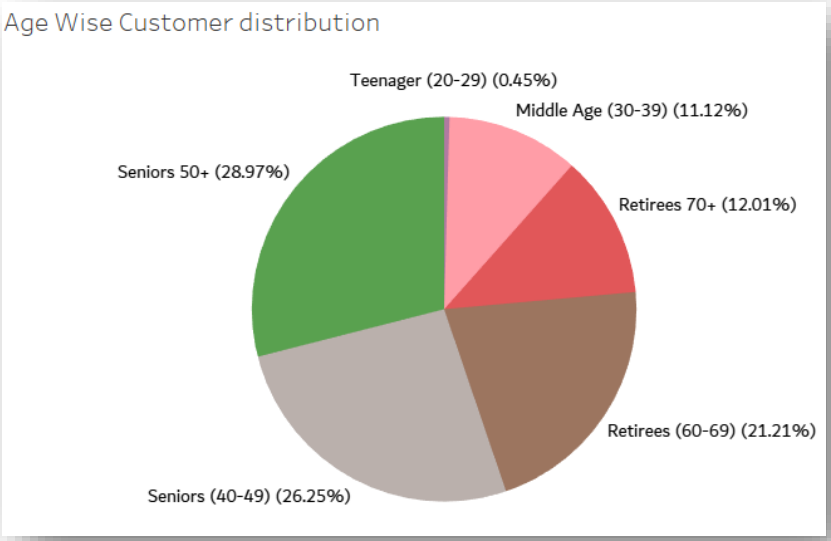
    - Age column is calculated from the “Year_Birth” column using “2023” as year or reference and binned as below for EDA purposes:

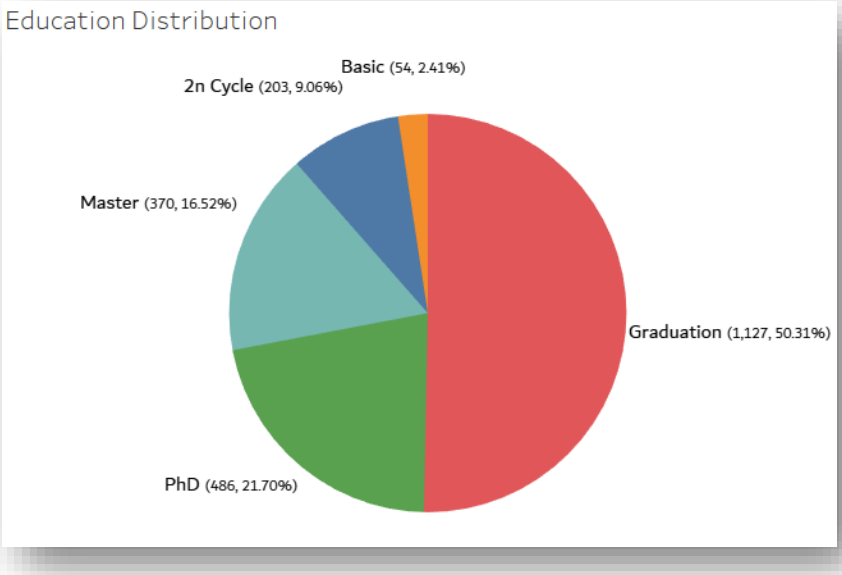
[Customer Age] >= 20 AND [Customer Age] < 30	Teenager (20-29)
[Customer Age] >= 30 AND [Customer Age] < 40	Middle Age (30-39)
[Customer Age] >= 40 AND [Customer Age] < 50	Seniors (40-49)
[Customer Age] >= 50 AND [Customer Age] < 60	Seniors 50+
[Customer Age] >= 60 AND [Customer Age] < 70	Retirees (60-69)
[Customer Age] >=70	Retirees 70+

Exploratory Data Analysis (EDA)

As part of the Exploratory Data Analysis, Tableau has been used for generation of visualization across the difference categorical columns against their distribution vectors.

Data bins have been created across Age and Income columns in order to reduce the complexity in the representation.

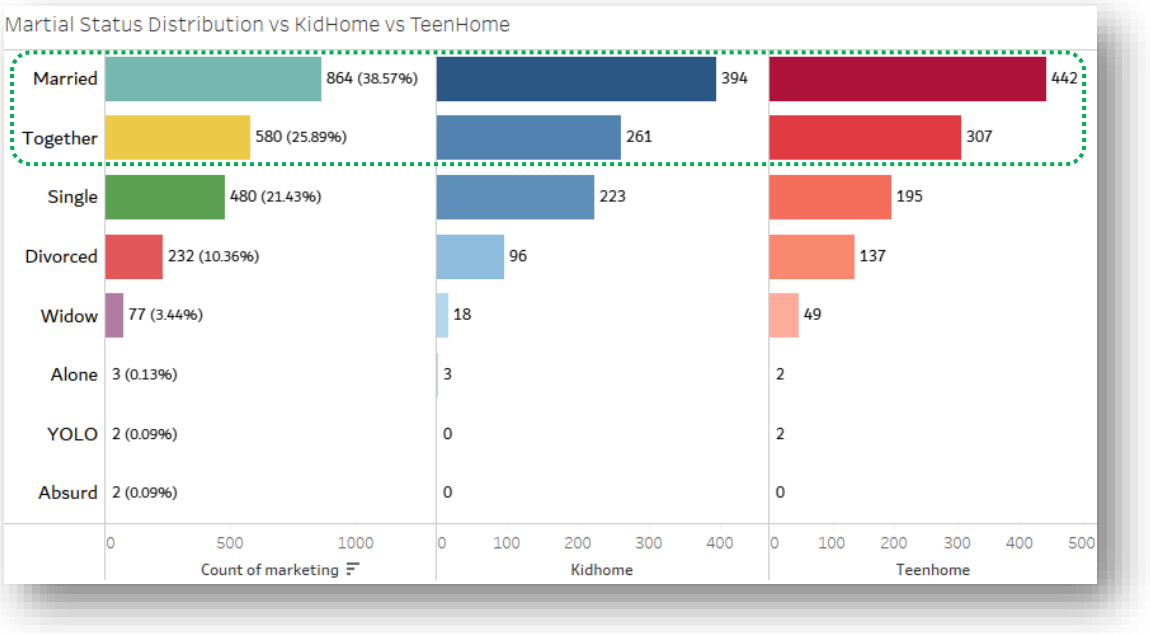




We observe that

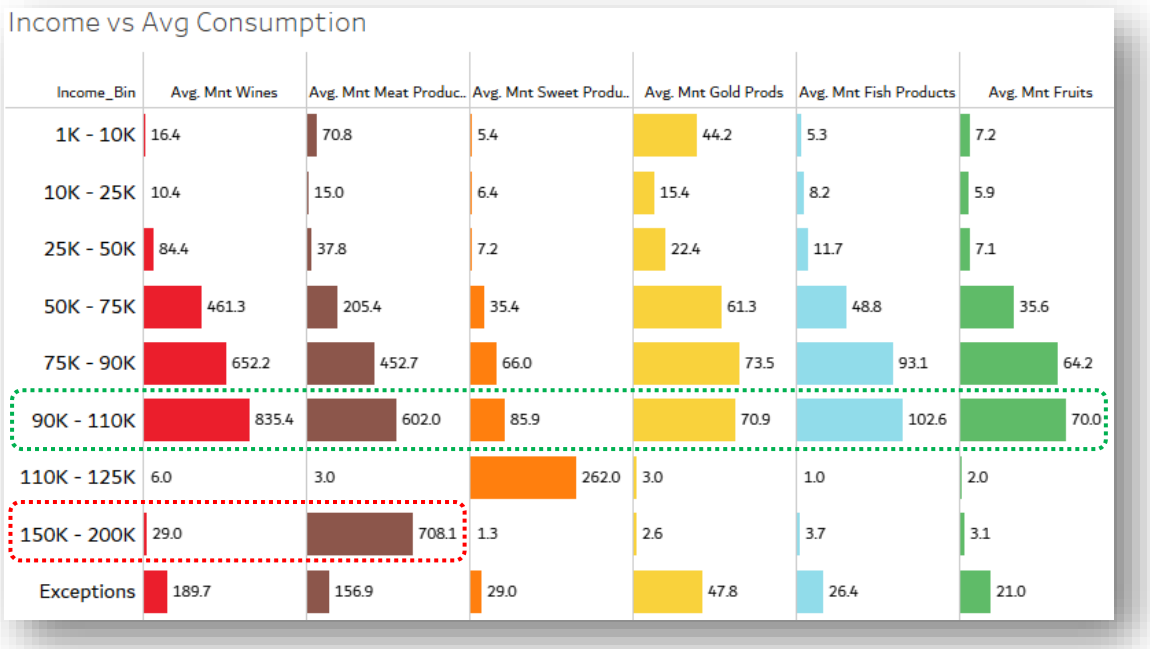
- Highest number of customers under analysis belong to Seniors who are aged between (50 – 60 yrs.) followed by Seniors in age group of (40 – 50 yrs.)
- The customer base has ~50% customers who are graduates followed by PhDs (~22%).

- The highest percentage of customer are either “Married (~39%)” or “Live Together (~26%)”. Majority of the Married/Living Together customers have Kid or Teenagers at home.

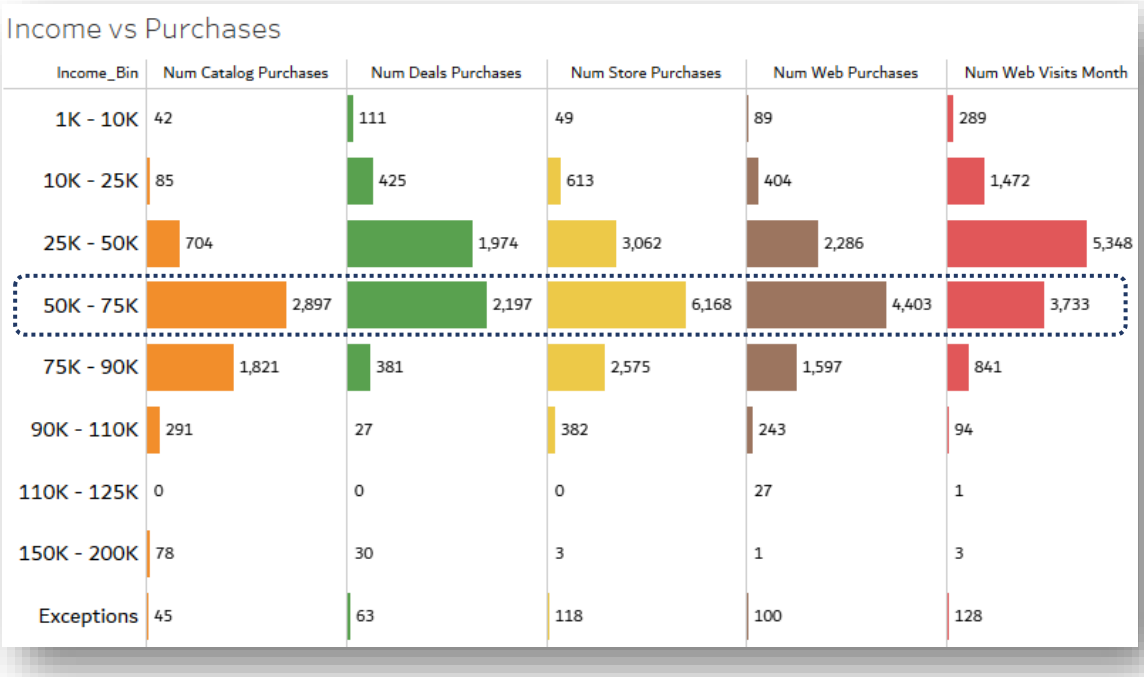


- On comparison of avg. monthly consumption (wine, meat, sweet, gold, fish products) against income ranges, we can see that high consumption observed among customers having income in range 90K – 110K in all product categories.

- Avg. monthly meat product consumption is highest among customers with income in range 150K – 200K.

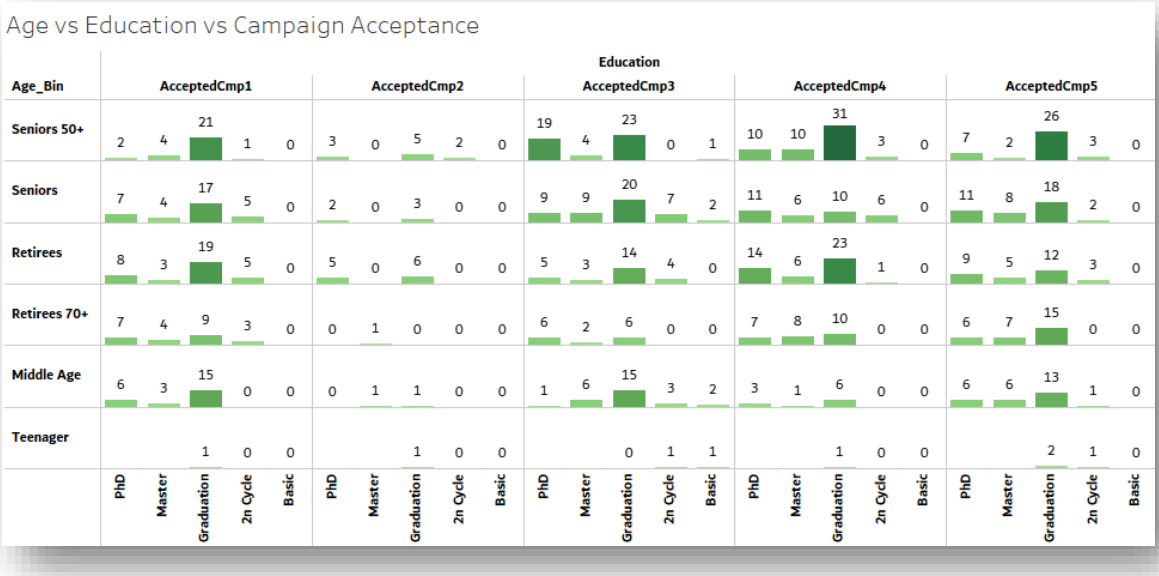


- When comparing income bins against the number of purchases across different categories (Catalog, Deals, Store, Web, Web visits), we observe the highest contribution among customers with income in range of 50K – 75K.



- To understand the impact of campaigns conducted, we perform a cross-section analysis of the acceptance against age and education levels. This analysis indicates higher acceptance ratio among graduate Seniors (50 -60 yrs.) and Retirees (60 – 70 yrs.).

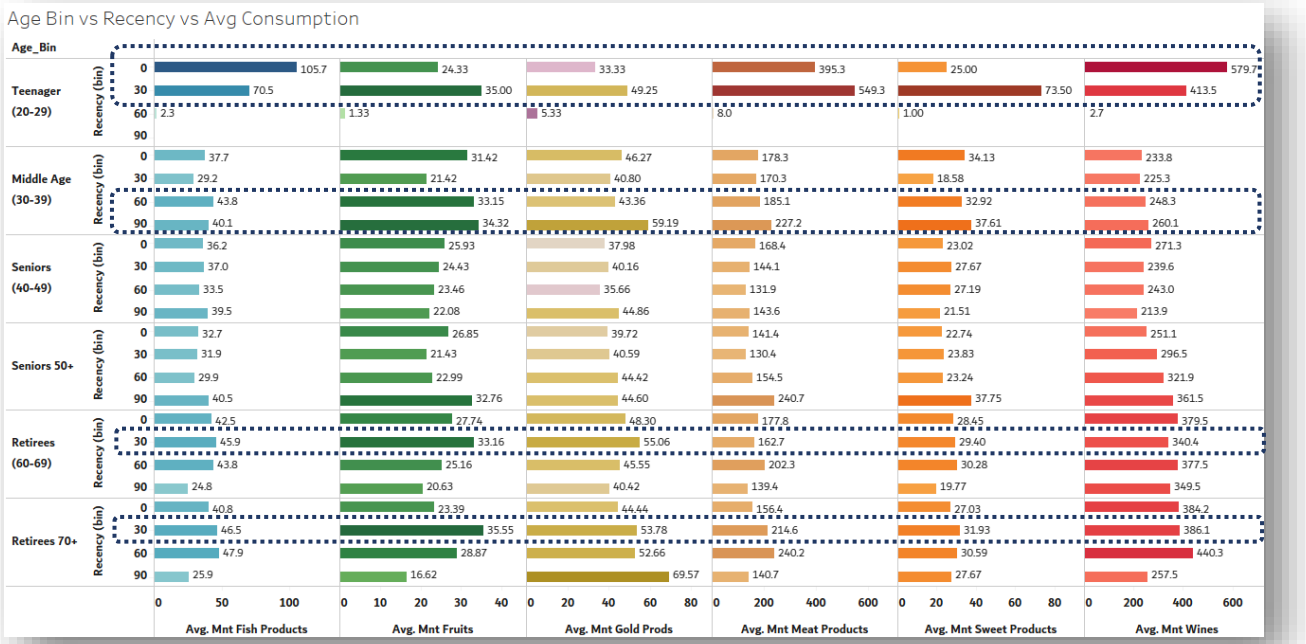
- Basic and 2nd cycle education level customers have lowest acceptance ratio.



- The comparison of average consumption across different age groups against recency feature (30-day bins) shows that Teenagers (20-29 yrs.) are more recent on their purchases at the stores along with high average spending on Wines and Fish Products.

- Middle Age (30-39 yrs.) customers are more of bi-monthly or tri-monthly buyers.

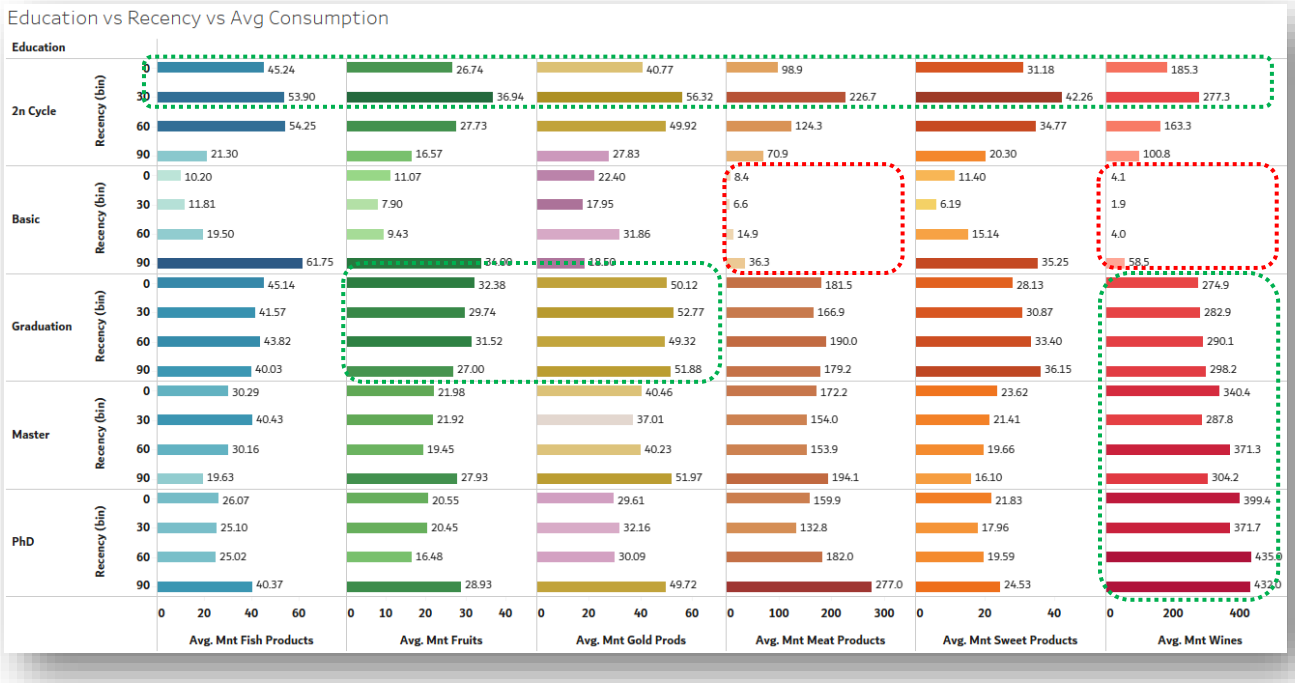
- Retirees (60-69 yrs.) and Retirees (70+) customer are more of monthly buyers.



- The comparison of average consumption across different education levels against recency feature (30-day bins) shows that Basic education level customers have the lowest consumption of Wines and Meat products. Graduation, Master and PhD customers however have a higher avg. consumption of Wines on the recency range.

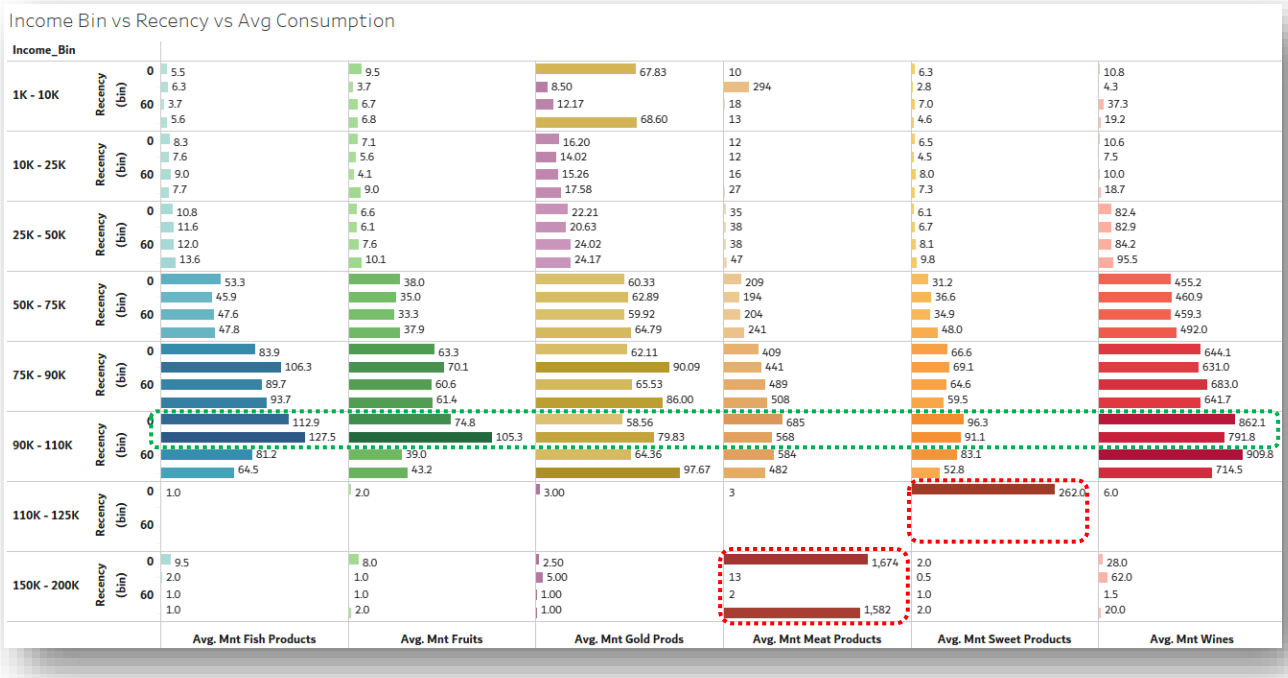
- Graduate customers show a higher avg. consumption of gold products and fruit products across different recency bins.

- 2nd Cycle Education customers show the highest avg. consumption across all the product categories on a 30-day/monthly cycle.



- The comparison of average consumption across different income brackets against recency feature (30-day bins) shows that customers with income in 90K-110K range have the highest avg. consumption across products on 30-day monthly cycle.

- Strangely, the customers in income range of 110K-125K are highest frequency buyers of sweet products. Similar anomaly is observed for customers in income range 150K-200K wherein the majorly buy Meat products.



- A key observation when comparing the median incomes across income bins with respect to the total purchases, we can see that the high-income group show lower purchases compared to 25K to 90K income customers. The retailer needs to understand the reason for this lower purchase trend and how it can be improved via. Promotional activities, introducing customer loyalty coupons or discount programs for higher purchase amounts.

Customer Spend Prediction

In order to predict customer spending, we will be processing the customer dataset to remove irrelevant columns, impute missing datapoints and perform necessary conversions in case of categorical features.

In the dataset, we found

- 24 rows with missing values/NaNs for Income column.
- Education and Martial_Status are categorical columns which can be one-hot encoded.
- Remove irrelevant columns (Z_CostContact, Z_Revenue, Response).

For the predication, we will be developing models based on below algorithms:

- Linear Regression (lm)
- Decision Tree (rpart)
- Random Forest (randomForest)
- XgBoost (xgboost)

Code for the prediction models

```
# Read data for retailer analysis
library(readxl)
setwd('D:/F/Learnings/IIMKAABI/R Material/R Dataset')
retail_data <- read_xlsx("W33836-XLS-ENG.xlsx", sheet = 'marketing')

# Remove unnecessary columns
retail_data <- retail_data[, -c(1, 27:29)]

# Calculate age of the customers using 2023 as year of reference
retail_data$Age <- 2023 - retail_data$Year_Birth

#Check for number of missing records
sum(is.na(retail_data))

# Using mice package for missing value imputation
library(mice)
retail_impute <- mice(retail_data, method = 'pmm', n=10, seed=1234)
retail_imputed_data <- complete(retail_impute)

retail_imputed_data$Dt_Customer_format <-
as.Date(as.numeric(as.character(retail_imputed_data$Dt_Customer)),origin = "1899-12-30")
retail_imputed_data$Dt_Customer_dt <- as.Date(retail_imputed_data$Dt_Customer, format = "%d-%M-%Y")
retail_imputed_data$Dt_Customer_updated <- as.Date(ifelse(is.na(retail_imputed_data$Dt_Customer_format),
                                    retail_imputed_data$Dt_Customer_dt,
                                    retail_imputed_data$Dt_Customer_format))

retail_imputed_data <- retail_imputed_data[, -c(1, 7, 27, 28)]

# Convert Education and Martial_Status column as One_hot encoded values
library(fastDummies)

retail_clean_data <- dummy_cols(retail_imputed_data, select_columns = c("Education", "Marital_Status"))
retail_clean_data <- retail_clean_data[, -c(1,2, 25)]
summary(retail_clean_data)
str(retail_clean_data)

# Clean column names
library(janitor)
retail_clean_data <- clean_names(retail_clean_data)
colnames(retail_clean_data)

# Define the list of target value list
target_var <-
c("mnt_wines","mnt_fruits","mnt_meat_products","mnt_fish_products","mnt_sweet_products","mnt_gold_prods")
```

```
retail_final <- retail_clean_data[!names(retail_clean_data) %in% target_var]

library(randomForest)
library(rpart)
library(xgboost)
library(Metrics)

# Define model run parameter hash
# install.packages('hash')
library(hash)


model_performance <- hash()
# Create the model list
models = c("lm","rpart","randomForest")
for (model_name in models){
  model_var_rmse <- hash()
  print(model_name)
  for (var in target_var){

    y <- retail_clean_data[var]
    X <- retail_final

    final_data <- cbind(X, y)
    # Split data into training and test sets
    set.seed(1234)

    index <- sample(1:nrow(final_data), .80*nrow(final_data))
    training <- final_data[index,]
    test <- final_data[-index,]

    # Train the model using training data
    options(scipen =10)
    m <- match.fun(model_name)
    if (model_name != "rpart") {
      model <- m(as.formula(paste0(var, "~ .")),data=training)
      predicted <- predict(model, newdata = test)
      model_rmse <- rmse(test[[var]], predicted)
    }
    else {
      model <- m(as.formula(paste0(var, "~ .")),data=training, method='anova', cp=0.01)
      predicted <- predict(model, newdata = test, type='vector')
      model_rmse <- rmse(test[[var]], predicted)
   }

    #1-mape(test[[var]], predicted)
    if (!has.key(var, model_var_rmse)) {
      model_var_rmse[var] <- model_rmse
    }  else {
      model_var_rmse[var] <- append(model_var_rmse[[var]], model_rmse)
    }

  }

  if (!has.key(model_name, model_performance)) {
    model_performance[model_name] <- model_var_rmse
  }  else {
    model_performance[model_name] <- append(model_performance[[model_name]], model_var_rmse)
  }
}

# Create xgboost model
model_name = "xgboost"
model_var_rmse <- hash()
print(model_name)
# XGboost modelling
for (var in target_var){

  y <- retail_clean_data[var]
  X <- retail_final

  final_data <- cbind(X, y)
  # Split data into training and test sets
  set.seed(1234)
```

```
  index <- sample(1:nrow(final_data), .80*nrow(final_data))
  training <- final_data[index,]
  test <- final_data[-index,]

  train_label <- as.numeric(training[,ncol(training)]) # dep variable of train data
  test_label <- as.numeric(test[,ncol(test)]) # dep variable of test

  train_features <- training[,1:ncol(training)-1]# train data's ivs
  test_features <- test[,1:ncol(test)-1]
  ## convert both test and train data-sets (IVs) into matrix

  train_features <- as.matrix(train_features)
  test_features <- as.matrix (test_features)
  #state the parameters
  parameters <- list(eta = 0.01,# learning rate
                     max_depth = 6, #depth of the tree
                     subsample = 1, # % sample used
                     colsample_bytree = 1, # # columns cused
                     min_child_weight = 1, #number of minimm instance in branch
                     gamma = 0, # auto pruning
                     eval_metric = "rmse", # what is the loss minimization function
                     booster = "gbtree") # it shows it is tree based algo

  #Running XGBoost
  model_xg <- xgboost(data = train_features,
                      label = train_label,
                      set.seed(1234),
                      nthread = 6, # how many cores we want to consume
                      nround = 5000,
                      params = parameters,
                      print_every_n = 500,
                      early_stopping_rounds = 20,
                      verbose = 1)

  predicted <- predict(model_xg, newdata = test_features)# predict the test data
  table1 <- data.frame(Actual  = test[,ncol(test)], Predicted = predicted)

  rmse_xg <- sqrt(mean((table1$Predicted-table1$Actual)^2))

  #importance drivers
  importance <- xgb.importance(feature_names = colnames(test_features),
                               model = model_xg)
  # save the plots
  plt_path1 = paste('D:/F/Learnings/IIMKAABI/R Material/Submission/Capstone/',var,'.jpeg')
  jpeg(file=plt_path1)
  xgb.plot.importance(importance_matrix = importance, top_n =  8)
  dev.off()

  #shap values
  plt_path2 = paste('D:/F/Learnings/IIMKAABI/R Material/Submission/Capstone/',var,'_shap.jpeg')
  jpeg(file=plt_path2)
  xgb.plot.shap(data = test_features,
                model = model_xg,
                top_n = 5)
  dev.off()

  if (!has.key(var, model_var_rmse)) {
    model_var_rmse[var] <- rmse_xg
  }  else {
    model_var_rmse[var] <- append(model_var_rmse[[var]], rmse_xg)
  }
}

if (!has.key(model_name, model_performance)) {
  model_performance[model_name] <- model_var_rmse
}  else {
  model_performance[model_name] <- append(model_performance[[model_name]], model_var_rmse)
}
```

Results

The prediction models comparison table based on Model vs Root Mean Square Error (RMSE):

Model Name	Mnt Wines	Mnt Fruits	Mnt MeatProducts	Mnt FishProducts	Mnt SweetProducts	Mnt GoldProds
Linear Regression	205.15	34.13	156.32	43.71	34.58	42.24
Decision Tree	200.79	36.4	144.66	45.82	35.92	42.84
xgBoost	156.86	31.83	128.48	43.71	34.41	44.8
Random Forest	152.8	30.92	127.29	40.28	32.55	41.13

We can observe that without any hyper tuning performed on the model, the basic RandomForest algorithm performs better in terms of lowest RMSE values amongst the 4 models considered.

Note: *The models haven't been fine-tuned using hyper parameters. Hyperparameter tuning could yield comparatively lower RMSE and better prediction models.*

Customer Segmentation

Considering the fact that the customer dataset has around 30 features, we will employ PCA (Principal Component Analysis) in order to reduce the dimensionality of the dataset thereby reducing multicollinearity within the data.

For performing PCA, we use the following packages in R:

- psych
- FactoMineR
- factoextra – For graphical visualization

Post PCA, we use the eigne values for find the number of factors which have eigen values greater than 1. These factors are then considered for segmentation using KMeans.

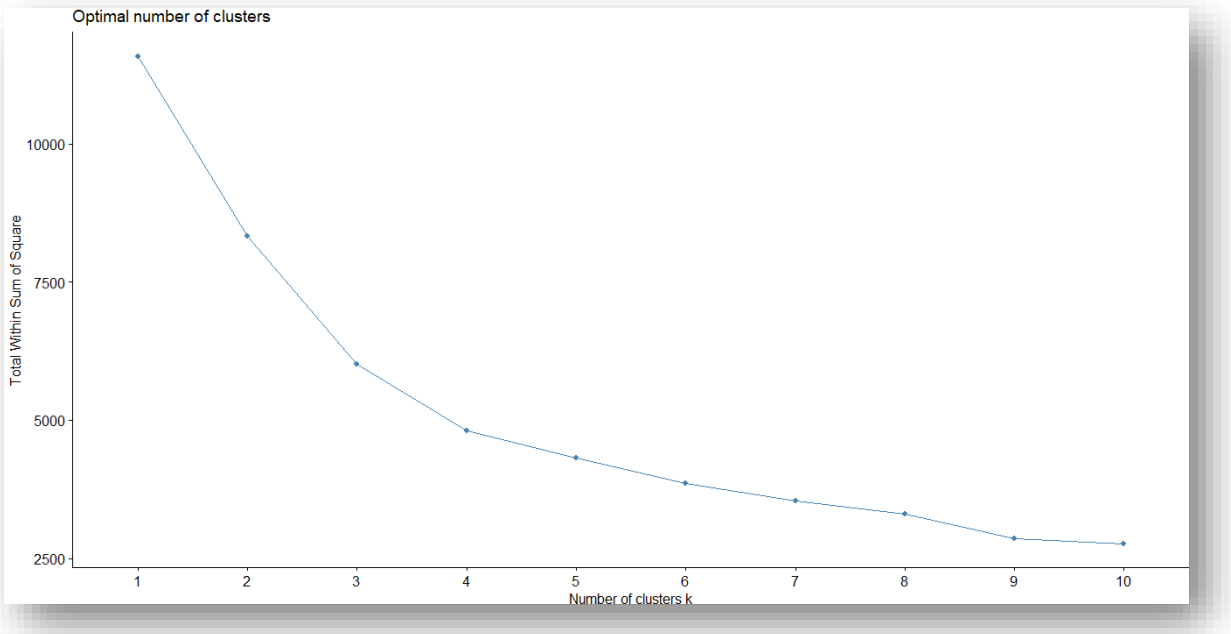
We observe that a cluster size of 4 helps in capturing the optimum customer segmentation or baskets with respect to the data under consideration.

Code for PCA and KMeans Segmentation

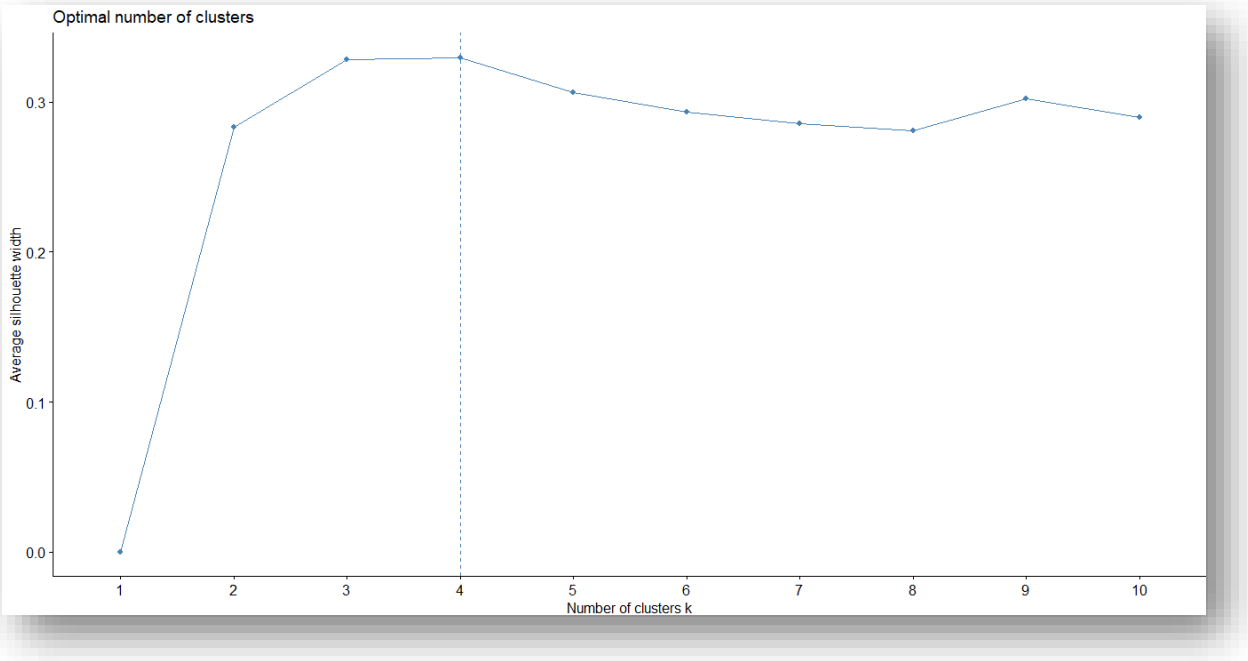
```
# READ THE DATA INTO R (W33836-XLS-ENG)
setwd('D:/F/Learnings/IIMKAABI/R Material/R Dataset')
library(readxl)

segment_data <- read_xlsx("W33836-XLS-ENG.xlsx", sheet = 'marketing')

# calculate the age of customer
library(dplyr)
segment_data <- segment_data %>% mutate(age= 2023 - Year_Birth)

pca_data <- segment_data[, -c(1:4, 8, 21:25, 26:29)]
pca_data <- na.omit(pca_data)

library(corrplot)

cor_matrix <- cor(pca_data)

corrplot(cor_matrix, method = "color", type = "upper",
         tl.col = "black", tl.srt = 45, tl.cex = 0.8, # Text color, rotation, and size
         number.cex = 0.6, # Coefficient size
         addCoef.col = "black", # Add coefficient of correlation
         col = colorRampPalette(c("blue", "white", "orange"))(200),
         cl.lim = c(0, 0),
         title = "Correlation Matrix of All Features",
         mar = c(1, 1, 1, 1)) # Margin for the title

# We observe there exists correlation between the variables which justifies the
# use of PCA for variable grouping.

#install.packages('psych')
#install.packages('FactoMineR')
#install.packages('factoextra')
library(psych)
library(FactoMineR)
library(factoextra)

# Perform PCA
pca_result <- PCA(pca_data, graph = FALSE)

# Visualize PCA results
fviz_pca_var(pca_result, col.var = "contrib", gradient.cols = c("black", "cyan", "blue", "red"),
             ggtheme = theme_minimal())

# Visualize the eigenvalues
fviz_eig(pca_result)

# Find the number of factors with eigenvalues greater than 1
```

```
# Extract eigenvalues
eigenvalues <- pca_result$eig


# Find the number of factors with eigenvalues greater than 1
num_factors <- sum(eigenvalues[, 1] > 1)
###
pca_psych <- principal(pca_data, nfactors = num_factors, rotate = "varimax")

score <- print(pca_psych$loadings)
factor_scores <- pca_psych$scores

factor_scores <- data.frame(factor_scores)

names(factor_scores)[1] <- 'income_and_amount_spending'
names(factor_scores)[2] <- 'deals_and_visits'
names(factor_scores)[3] <- 'age_and_teenhome'
names(factor_scores)[4] <- 'recency'


#### cluster analysis using factor_score
#### start with k=4, using 4 clusters

set.seed(1234)
cluster_no_4 <- kmeans(factor_scores, 4)

cluster_no_4$tot.withinss
# > cluster_no_4$tot.withinss
# [1] 4811.094

factor_scores$cluster_four <- cluster_no_4$cluster
#install.packages("factoextra") for elbow plot
library(factoextra)
fviz_nbclust(factor_scores, kmeans, method='wss') # For elbow plot

# Create Silhouette score graph
fviz_nbclust(factor_scores, kmeans, method='silhouette')
```

Elbow Plot:



Silhouette Score:



Insights and Future Actions

From the overall analysis performed on the limited customer data considered, below are the insights and actions for the retailer.

- **Focused Customer Initiatives:**

As per the EDA, we have observed areas which needs focus from the retailer in terms on improving customer experience and increasing the customer spending culture. This could be by means of introducing store loyalty cards or membership discount for customers who belong to high income and low spending groups.

There needs to be focused campaigning activity in order to capture the low spending customer groups and to understand the areas which needs improved for better customer experience.

Personalized Customer Experiences: Leverage analytics to tailor recommendations and promotions based on individual preferences. Personalization enhances customer satisfaction and loyalty.

Sustainability: Embrace eco-friendly practices. Offer reusable bags, reduce plastic packaging, and promote locally sourced, organic products. Consumers increasingly value sustainability.

Home Delivery: Invest in efficient home delivery services. As online shopping grows, convenient and reliable delivery options are essential.

Hybrid Shop-and-Eat Concepts: Consider integrating dining areas or cafes within your supermarket. This allows customers to shop and enjoy meals in one convenient location.

- **Customer Engagement:**

There are customer groups which currently don't tend to be engaged with the store products or aren't spending in accordance to their income capacity. The retailer needs to conduct surveys or capture post sales feedbacks to understand the overall customer experience:

  - Did you get what you were looking for?
  - How easy was the product placement?
  - How helpful were the store staff with the shopping experience?
  - What could improve with respect to product quality, product inventory and checkout experience?

Similar questionnaires or feedbacks can help in understanding the issues that the customers might be facing which led to lower spending in certain product or customer groups.

In-Store Experiences and Events: Encourage customers to spend more time in your stores by hosting pop-ups, events, and classes. Create engaging experiences that keep shoppers coming back.

Aiding Product Discovery: Use technology to help customers discover new products. Interactive displays, QR codes, and mobile apps can guide shoppers to relevant items.

- **Cross-selling and Upselling:**

The retailer needs to identify areas and products which will be more effective in terms of customer satisfaction and employ cross-selling or upselling strategies to high income customer groups. This would help in increasing the revenue/ROI for the retailer and help in conversion of frequent/to-be-loyal customers with minimum capital investment.

- **Churn Identification:**

Exhaustive analysis to identify the high churn risk customer groups especially among high income and frequent customer groups and creating tailormade catalogues of products or discount programs for such risky customer groups.

Conclusion

The data analysis project analysed customer behaviour using Tableau, R and Machine Learning algorithms resulted in the below conclusions which could come in as supporting aid for the retailer when deciding future actions.

Customer Segmentation: Analyse customer demographics, behaviour, and preferences. Segmentation helps tailor marketing efforts and improve customer satisfaction.

Product Performance: Evaluate the performance of individual products. Which items sell well? Which need improvement or promotion?

Pricing Insights: Understand price elasticity and optimal pricing points. Adjust prices based on demand and competition.

Promotion Effectiveness: Assess the impact of promotions, discounts, and loyalty programs on sales and customer retention.

Recommendations

Since the current data under the scope of analysis was limited, it would be helpful to include supply chain details and sales trends which can help in enhancing the present analysis done.

Supply Chain Optimization: Optimize supply chain processes, including inventory replenishment, logistics, and supplier relationships.

Sales Trends: Identify patterns in sales data, such as peak shopping hours, popular product categories, and seasonal variations. This information can guide inventory management and marketing strategies.

Advanced Modelling Algorithms: Use of Deep learning or advanced machine learning models can sizably improve the prediction capability of the underlying models. This will be vital for future analysis activities.

Additionally, the below challenges need to be also addressed to improve the overall journey of data analysis.

Data Gathering: Extracting meaningful insights from consumer information can be challenging. Retailers need to sift through vast amounts of data to identify actionable patterns and trends. Appropriate data capture mechanisms have to be implemented to further collate the source information.

Privacy Concerns: Retailers collect both general demographic statistics and consumer-specific details. Balancing data utilization with privacy compliance is crucial.

Data Quantity: Managing large volumes of data is a significant challenge. Retailers must handle data from various sources, including point-of-sale systems, loyalty programs, and social media.

Data Quality: Ensuring accurate and reliable data is essential. Inaccuracies can lead to flawed analyses and misguided decisions.

Standardization: Data may come in different formats, making it challenging to integrate and analyse consistently. Standardizing data structures is vital for meaningful insights.