

# Project Proposal : Group 7

## **Project name:**

Customer profiling, store design and revenue maximization for Food Mart Retailer

## **Executive Summary:**

We live in a world where there are huge monopolies in the retail sector (Walmart and Amazon) Even in the face of such staunch competition, other franchises such as Aldi, Dollar General and 7-Eleven have successfully survived. The two main parameters which any retail store focuses on is overall revenue, and number of customers. In this project, we aim to study and analyze data of Convenient Food Mart (CFM), founded in Chicago in 1958. First, we aim to do a thorough analysis on the dataset. We would like to understand consumer patterns (how often a consumer buys Bread along with milk) and try presenting analysis on the item stocking recommendation for the mart using Apriori Algorithm, as well as the cost borne by CFM for acquiring a new customer (via advertisement).

## **About Food Mart:**

CFM is a chain of convenience stores in the United States. The private company's headquarters are located in Mentor, Ohio, and there are currently approximately 325 stores located in the US. Convenient Food Mart operates on the franchise system.

Convenient Food Mart was the nation's third-largest chain of convenience stores as of 1988. The NASDAQ exchange dropped Convenient Food Mart the same year when the company failed to meet financial reporting requirements.

Source - [https://en.wikipedia.org/wiki/Convenient\\_Food\\_Mart](https://en.wikipedia.org/wiki/Convenient_Food_Mart)

## **Dataset Information:**

The raw dataset for this is available at github - <https://github.com/jpvelez/foodmart>

The raw dataset is divided into 8 distinct datasets with a total of about 200K data points with over 75 features. We were able to find a refined version of the same dataset at kaggle - <https://www.kaggle.com/datasets/ramjasmaurya/medias-cost-prediction-in-foodmart>.

- **Customer Profile:** Marital Status, gender, total children, occupation, annual income, education level, membership status, etc.
- **Store Information:** Store sales, costs, building type, market area, meat/grocery/frozen food space, city/state location, etc.
- **Campaign Information:** Sales country, promotion name, customer acquisition cost, etc.
- **Product data:** Brand, unit sales, weight, product type, food type, etc.

## **Project Details: Who will be our audience?**

The demographics that will benefit most from the analysis performed would be:

- Food Mart Management - for revenue maximization
- Store Design Team - to design store in the most efficient way
- Food Mart Procurement team - to decide which items to order with what frequency

While our project is relatively on a smaller scale, it can conceptually provide a broader perspective of marketing and customer retention for the aforementioned audiences. The observation of customer trends could generate ideas not only for Food Mart, but potentially other supermarket management teams in order to create cost-efficient marketing campaigns that are more appealing and catered towards various preferences. The inferences derived from the product sales could also inspire changes in store layout to prioritize popular products and shine a wider spotlight on them. With customer profiling being an integral aspect of our analysis, Food Mart could use customer information to check different advertising methods for increasing revenue.. Finally, we hope that our project could enrich our colleagues' understanding of the numerous business scenarios in which one could apply their analytical skills and knowledge of data science.

## **What are your research question(s)?**

- (i) Finding the most cost effective customer acquisition method.
- (ii) Which city/country has the highest sales in which department( or category of food)?

- (iii) What kind of promotions persuade what kind of demographics (Relation between advertisement and customer profile)
- (iv) Do factors like level of education, marital status, family size, and income influence customer product preferences?
- (v) Determine which city has the highest spending and why is it the highest as well?

## **What predictive model(s) are you planning to run?**

We have multiple suggestions in the team, but we might choose to change the models listed below:

**(i) Regression Model:** Linear Regression Model

**(ii) Decision Tree Models:** Random Forest, XGBoost Model

$X \rightarrow$  All relevant features decided after PCA Analysis and Feature Correlation

### **a. What is the importance of your project**

- To organize a better company strategy for a retailer. Recognize your profit margins and work to them while keeping in mind the cost of client acquisition.
- Customer acquisition is the main component in business, so it's imperative to know which form of advertisement suits which type of customers, instead of spending on low yielding channels.
- The project could provide a deeper perspective into possible correlations with customer lifestyle and preferences, which in turn could aid in yielding more engaging and impactful marketing campaigns for specific types of customers.
- The project will provide us a thorough examination of the expenditure in different segments of the market, providing investors with better suggestions on where to make cost reductions and increase yield.

### 3. Provide a code that: We have attached an R file along with the submission. Please find the code in that. {More explanation given in the R file}

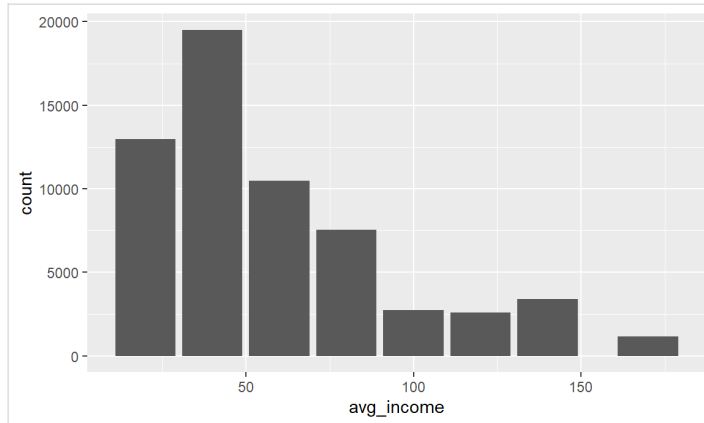
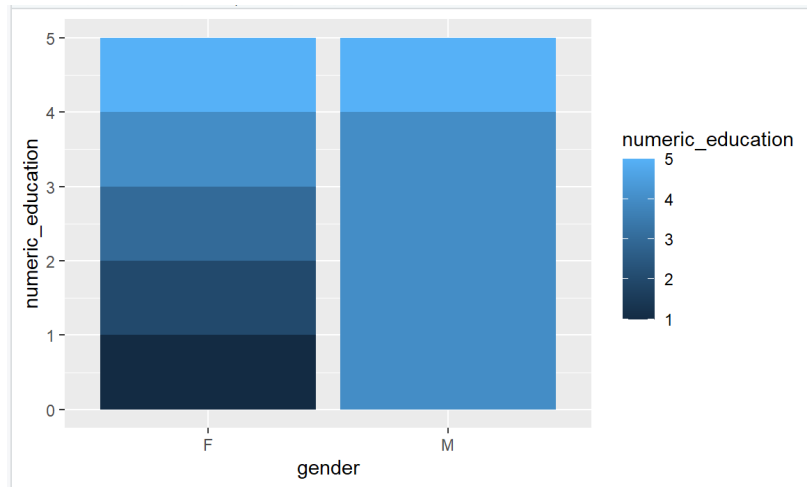
b. Imports your data :

Data	
FoodmartData	60428 obs. of 40 variables
\$ food_category	: chr "Breakfast Foods" "Breakfast Foods" "Break...
\$ food_department	: chr "Frozen Foods" "Frozen Foods" "Frozen Food...
\$ food_family	: chr "Food" "Food" "Food" "Food" ...
\$ store_sales.in.millions.	: num 7.36 5.52 3.68 3.68 4.08 4.08 5.44 3.74 4...
\$ store_cost.in.millions.	: num 2.72 2.59 1.36 1.18 1.43 ...
\$ unit_sales.in.millions.	: num 4 3 2 2 3 3 4 2 3 3 ...
\$ promotion_name	: chr "Bag Stuffers" "Cash Register Lottery" "Hi...
\$ sales_country	: chr "USA" "USA" "USA" "USA" ...
\$ marital_status	: chr "M" "M" "S" "M" ...
\$ gender	: chr "F" "M" "F" "F" ...
\$ total_children	: num 1 0 4 2 0 2 4 1 2 3 ...
\$ education	: chr "Partial High School" "Bachelors Degree" "...
\$ member_card	: chr "Normal" "Silver" "Normal" "Bronze" ...
\$ occupation	: chr "Skilled Manual" "Professional" "Manual" ...
\$ houseowner	: chr "Y" "Y" "N" "Y" ...
\$ avg_cars_at.home.approx.	: num 1 4 1 2 2 1 2 4 2 1 ...
\$ avg..yearly_income	: chr "\$10K - \$30K" "\$50K - \$70K" "\$10K - \$30K" ...
\$ num_children_at_home	: num 1 0 0 2 0 2 0 0 0 ...
\$ avg_cars_at.home.approx..1	: num 1 4 1 2 2 1 2 4 2 1 ...
\$ brand_name	: chr "Carrington" "Carrington" "Carrington" "Ca...
\$ SRP	: num 1.84 1.84 1.84 1.84 1.36 1.36 1.36 1.87 1...
\$ gross_weight	: num 19.7 19.7 19.7 19.7 7.12 7.12 7.12 16.7 7...
\$ net_weight	: num 17.7 17.7 17.7 17.7 5.11 5.11 5.11 14.7 5...
\$ recyclable_package	: num 1 1 1 0 0 0 1 0 1 ...
\$ low_fat	: num 0 0 0 0 1 1 1 1 0 ...
\$ units_per_case	: num 17 17 17 17 29 29 29 10 29 25 ...
\$ store_type	: chr "Deluxe Supermarket" "Deluxe Supermarket" ...
\$ store_city	: chr "Salem" "Salem" "Salem" "Salem" ...
\$ store_state	: chr "OR" "OR" "OR" "OR" ...
\$ store_sqft	: num 27694 27694 27694 27694 27694 ...
\$ grocery_sqft	: num 18670 18670 18670 18670 18670 ...
\$ frozen_sqft	: num 5415 5415 5415 5415 5415 ...
\$ meat_sqft	: num 3610 3610 3610 3610 3610 3610 3610 3610 36...
\$ coffee_bar	: num 1 1 1 1 1 1 1 1 1 ...
\$ video_store	: num 1 1 1 1 1 1 1 1 1 ...
\$ salad_bar	: num 1 1 1 1 1 1 1 1 1 ...
\$ prepared_food	: num 1 1 1 1 1 1 1 1 1 ...
\$ flav...	: num 1 1 1 1 1 1 1 1 1 ...

c. Tidy your data (including data transformations) : Here we have used tidy data concepts, and transformed the “avg..yearly\_income” from a categorical variable to a numerical value and further splitted the data into minimum, maximum and average Income to ease our EDA process and enhance other analysis.

customer_informa...	
customer_informa...	60428 obs. of 14 variables
\$ marital_status	: chr "M" "M" "S" "M" ...
\$ gender	: chr "F" "M" "F" "F" ...
\$ total_children	: num 1 0 4 2 0 2 4 1 2 3 ...
\$ education	: chr "Partial High School" "Bachelo...
\$ member_card	: chr "Normal" "Silver" "Normal" "Br...
\$ occupation	: chr "Skilled Manual" "Professional...
\$ houseowner	: chr "Y" "Y" "N" "Y" ...
\$ avg..yearly_income	: chr "\$10K - \$30K" "\$50K - \$70K" "\$...
\$ total_children.1	: num 1 0 4 2 0 2 4 1 2 3 ...
\$ avg_cars_at.home.approx.	: num 1 4 1 2 2 1 2 4 2 1 ...
\$ avg_income	: num 20 60 20 40 40 60 40 60 20 80 ...
\$ minimum_income	: num 10 50 10 30 30 50 30 50 10 70 ...
\$ maximum_income	: num 30 70 30 50 50 70 50 70 30 90 ...
\$ numeric_education	: num 1 4 1 2 1 4 2 1 1 5 ...

- d. Conduct a simple Exploratory Data Analysis [Only few images have been attached]



- e. Provide Descriptive Analytics:

In our R file, we have provided a descriptive analysis. Including a step-by-step explanation of why we conducted the descriptive analysis and what inference we are trying to obtain.