# Cost Prediction on Acquiring Customers In Marketing campaign

**Project done by:**

**Ganesh Sundar**

# Table of Contents

# Industry Review

**1. The Digital Transformation**
**E-commerce Surge:** Online shopping has seen exponential growth over the last decade. Platforms like Amazon, Alibaba, and Shopify have become household names, offering consumers unparalleled convenience and variety. The COVID-19 pandemic further accelerated this trend, with many consumers preferring online shopping over traditional brick-and-mortar stores for safety reasons.

**2. Consumer-Centric Approaches**
**Personalization:** With the aid of data analytics and AI, retailers can now offer personalized shopping experiences. From product recommendations to tailored marketing messages, personalization has proven to increase loyalty and sales.
**Sustainability:** Modern consumers are more environmentally conscious. Brands like Patagonia and H&M are responding by offering sustainable products and showcasing eco-friendly practices, aligning with consumers' values and driving brand loyalty**.**

**3. Supply Chain Innovations**
**Automation and Robotics:** Warehouses are becoming more automated. Companies like Ocado are using robots to pick and pack groceries, increasing efficiency and reducing human error.
**Blockchain:** Transparency in product sourcing is becoming vital. Blockchain allows for traceability, ensuring products are ethically sourced and genuine. This is particularly significant in industries like luxury goods and pharmaceuticals.

**4. Emerging Markets**
**Asia's Dominance:** Asia, especially China and India, has seen rapid retail growth. The expanding middle class, combined with tech-savvy younger generations, has made these markets lucrative for both local and international retailers.
**Localized Strategies:** Global brands are realizing that one size doesn't fit all. Companies are adapting their products and marketing strategies to cater to local tastes and preferences, ensuring better market penetration.

**5. The Future of Retail**
**Virtual and Augmented Reality (VR & AR):** Virtual try-ons, immersive shopping experiences, and augmented store navigation are just a few ways VR and AR are set to revolutionize retail**.**
**Direct-to-Consumer (DTC):** Brands are increasingly bypassing retailers to sell directly to consumers. This approach offers higher profit margins and a direct relationship with the customer.
**Sustainable Practices:** As concerns about the environment grow, retailers will need to adopt sustainable practices not just in products but also in packaging, logistics, and more.

# Current Practices

**Multi-channel Marketing**

- Omnichannel Strategy: Ensuring a seamless user experience across all channels, whether online or offline.
- Cross-channel Analytics: Measuring the effectiveness of a campaign across different platforms to optimize ROI.

**Content Marketing**

- Value-driven Content: Providing informative, entertaining, or educational content that adds value for the audience.
- SEO Practices: Optimizing content for search engines to increase visibility and organic traffic.

**Social Media Advertising**

- Targeted Ads: Using platforms like Facebook and Instagram to target specific audience groups based on interests, behaviours or demographics.
- Influencer Collaborations: Partnering with influencers to tap into their audience and enhance brand credibility.

**Retargeting and Remarketing**

- Pixel-based Retargeting: Serving ads to users who've visited the website but didn't convert.
- List-based Remarketing: Targeting users from a specific email or contact list with tailored offers or content.

**Video Marketing**

- Engaging Video Content: Using platforms like YouTube or TikTok to share brand stories, product demos, or customer testimonials.
- Video Ads: Running short video advertisements on platforms like YouTube, Facebook, or Instagram.

**Loyalty Programs**

- Rewards Systems: Offering points, discounts, or exclusive offers to loyal customers to encourage repeat purchases.
- Referral Programs: Encouraging current customers to refer friends or family in exchange for rewards.

**Affiliate Marketing**

- Partnership Programs: Collaborating with affiliates to promote products/services in exchange for a commission on sales generated from their referrals.

**Reference**: https://link.springer.com/article/10.1007/s11747-022-00839-w

# Dataset and Domain

**Python Version:**

'3.10.9 | packaged by Anaconda, Inc. | (main, Mar 1 2023, 18:18:15) [MSC v.1916 64 bit (AMD64)]'

**Dataset Reference:** https://www.kaggle.com/datasets/ramjasmaurya/medias-cost-prediction-in-foodmart

# Literature Survey

CAC is a critical marketing metric for businesses to understand their marketing strategies and customer growth. This is the cost spent by organization to earn the customer. Many companies especially e-commerce-based companies who provides business to customer (B2C) are aggressive to earn a customer in what analysts have called a “land grab”. The motive of these early investments in customer acquisition is to gain the customer loyalty to stream the profits (Pei-Yu Chen, Lorin M. Hitt 2000). Many consumers experience non-negligible costs when switching between brands of services or products. According to Klemperer (1989), there are three types of switching costs: learning cost, transaction costs, and artificial costs. These switching cost called as a ‘Brand loyalty’ in marketing literature. Customer acquisition is a time-consuming and costly process. It is easier to attract existing customers than to new customers with a high attrition rate (Reinartz & Kumar, 2003). However, only a less proportion of customers become profitable to the company.

# Dataset and Domain

**Python Version:**

'3.10.9 | packaged by Anaconda, Inc. | (main, Mar 1 2023, 18:18:15) [MSC v.1916 64 bit (AMD64)]'

**Dataset Reference:** https://www.kaggle.com/datasets/ramjasmaurya/medias-cost-prediction-in-foodmart

## Data Dictionary form a table:

| Sl. No | Column name | Description | Data type |
|---|---|---|---|
| 1 | food_category | Category of the food item. | object |
| 2 | food_department | Department where the food item belongs. | object |
| 3 | food_family | Family classification of the food item. | object |
| 4 | store_sales(in millions) | Sales of the store in millions. | float64 |
| 5 | store_cost(in millions) | Cost incurred by the store in millions. | float64 |
| 6 | unit_sales(in millions) | Unit sales in millions. | float64 |
| 7 | promotion_name | Name of the promotion. | object |
| 8 | sales_country | Country where the sales occurred. | object |
| 9 | marital_status | Marital status of the customer. | object |
| 10 | gender | Gender of the customer. | object |
| 11 | total_children | Number of children. | float64 |
| 12 | education | Education Qualification | object |
| 13 | member_card | Which category of member the customer is | object |
| 14 | occupation | Occupation details | object |
| 15 | houseowner | Are they houseowner | object |
| 16 | avg_cars_at home(appro | Average Number of cars at home | float64 |
| 17 | avg. yearly_income | Average Yearly Income | object |
| 18 | num_children_at_home | Number of children at Home | float64 |
| 19 | brand_name | product brand name. | object |
| 20 | SRP | Selling Retail price of Product | float64 |
| 21 | gross_weight | gross weight of Product | float64 |
| 22 | net_weight | net weight of Product | float64 |
| 23 | recyclable_package | Is Product recyclable_package | boolean |
| 24 | low_fat | Is this a Low fat product | boolean |
| 25 | units_per_case | Units_per_case of Product | float64 |
| 26 | store_type | Type of store | object |
| 27 | store_city | City where store is present | object |
| 28 | store_state | State where store is present | object |
| 29 | store_sqft | Square feet of store. | float64 |
| 30 | grocery_sqft | grocery square feet of store | float64 |
| 31 | frozen_sqft | frozen square feet of store | float64 |
| 32 | meat_sqft | meat square feet of store | float64 |
| 33 | coffee_bar | Coffee bar present in store | boolean |
| 34 | video_store | Video store present in store | boolean |
| 35 | salad_bar | Salad bar present in store | boolean |
| 36 | prepared_food | Prepared food present in store | boolean |
| 37 | florist | Florist present in store | boolean |
| 38 | media_type | Media type present in store | object |
| 39 | cost | Cost required to acquire customer | float64 |

# Variable categorization (count of numeric and categorical columns)

| | |
|---|---|
| **Observation** | 60428 |
| **Variables** | 39 |
| **Number of Numerical Columns** | 22 |
| **Number of Categorical Columns** | 17 |

# Pre - Processing Data Analysis

**Missing Value:** There are 0 missing values
**Duplicated Rows:** There are 0 duplicated rows

As there is no missing value and duplicate rows, no imputation was done.

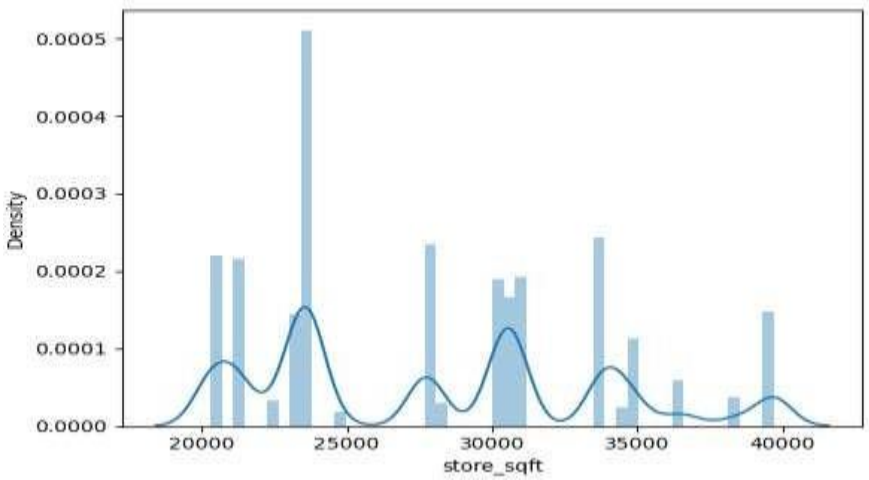
# Project Justification

**Project Statement:**

- Advertising and promotional efforts have become more regionalized in the business field.
- Retailers have become larger forcing marketers to shift money from advertising budgets to sales promotion.
- Marketers expect their promotional dollars to generate immediate sales but the cost spent on marketing activities like Mass media communication is in large amount for the acquisition of new cost which always does have the expected return
- Customer acquisition costs can cost up to 7 times more than selling to existing customers. Additionally, the probability of successfully selling to a new customer lies between 5 to 20 percent, whereas the probability of successfully selling to an existing customer lies between 60 to 70 percent

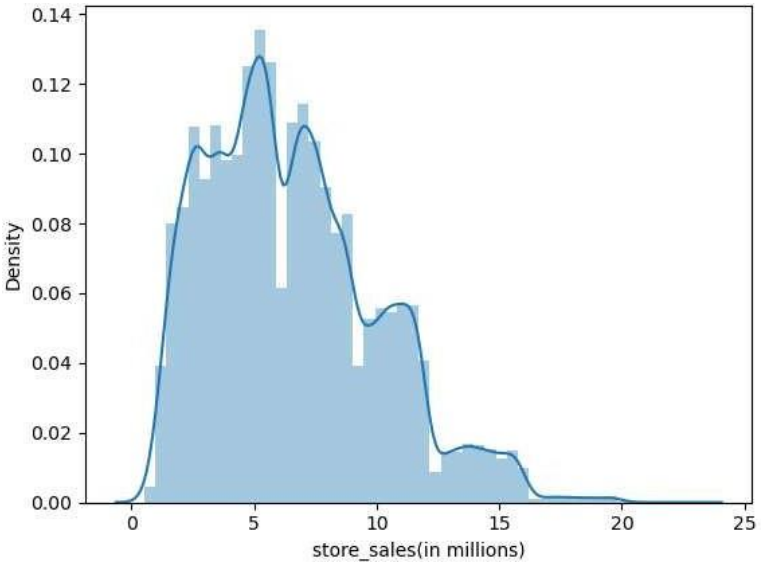
# Data Exploration (EDA)
## Univariate

❖ **Fig 1: Store square feet (Numeric Data Type)**



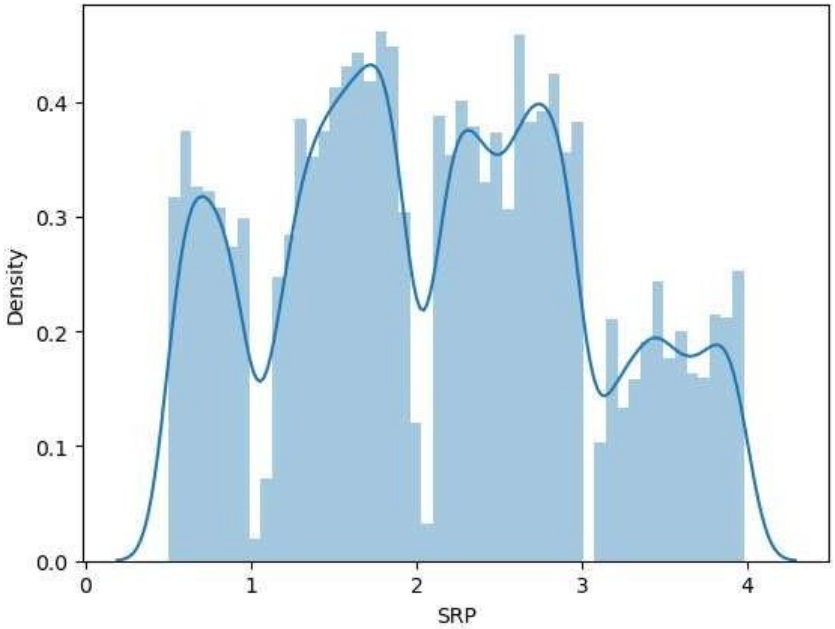

The total square feet of the shop where the marketing campaign is conducted is measured above
It is inferred that the shop ranges from 20k to 40k square feet and most the shop lies in 30k and 20.4k square feet.

❖ **Fig 2: Store Sales in millions (Numeric Data Type)**



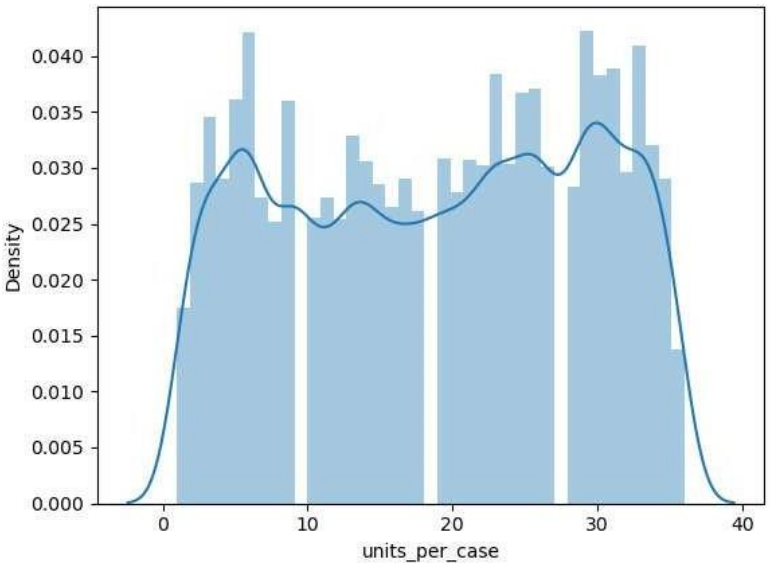

There is no normal distribution in the store sales . It is rightly skewed and most sales occurred between 5 to 6 million in each stores.

❖ **Fig 3: SRP (Selling retail price) (Numeric Data Type)**



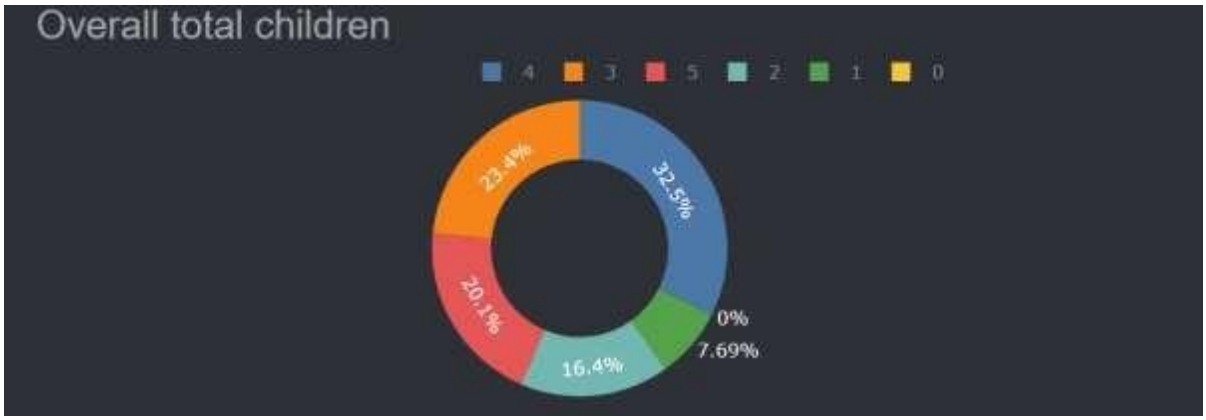

Selling retail price is normally distributed and it ranges from 0 to 4 from the distribution we can observe 3 bends where it consists of 4 group of clusters where each bend denotes each cluster

❖ **Fig 4: units per case (Numeric Data Type)**



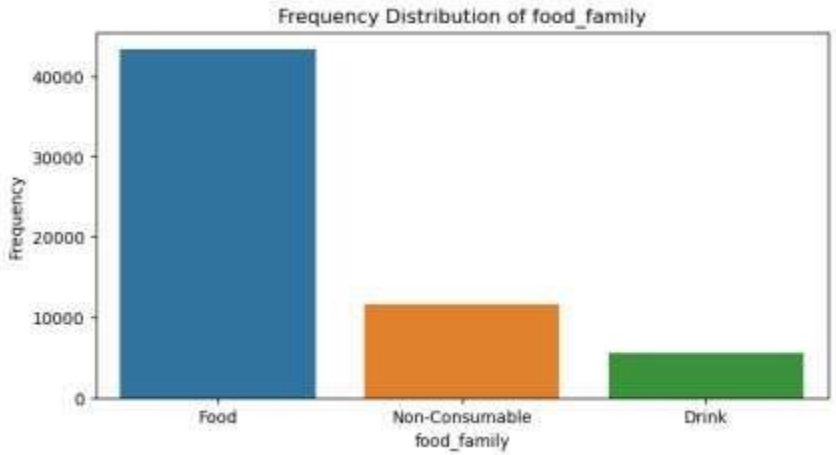

Unit per case seems to be normally distributed and this clearly forms 4 cluster and it ranges from 0 to 40

❖ **Fig5: Overall Total Children (Categorical Data Type)**




For the Marketing campaign, Family with 4 Children came to the campaigns mostly. Followed by 3 and 5 Children were inferred from the above pie chart.

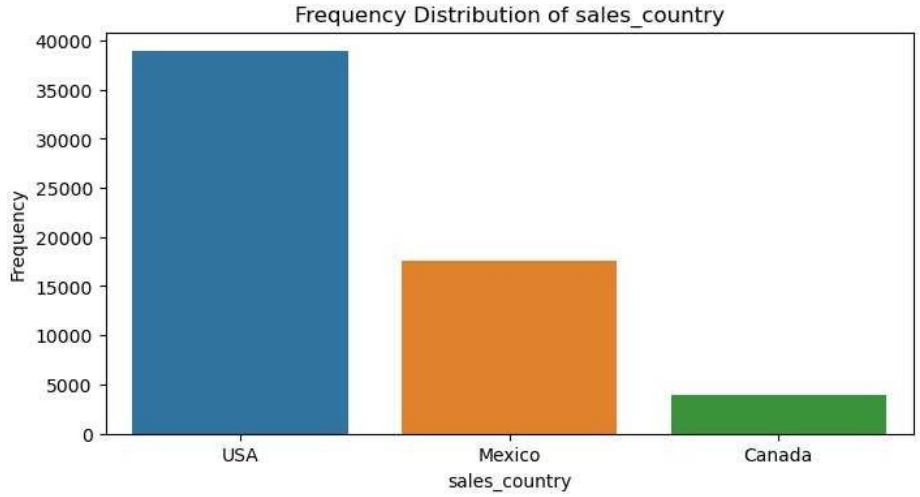
❖ **Fig 6 :Food Family (Categorical Data Type)**




The frequency distribution of the food family is displayed. It consists of 3 sub-groups. Food, Non-Consumable and Drink. The maximum frequency is under food category and most minimum consumed is Drink

| **Food Family** | **Frequency of Purchase** |
|---|---|
| Food | 44000 |
| Non-Consumable | 11000 |
| Drink | 6000 |

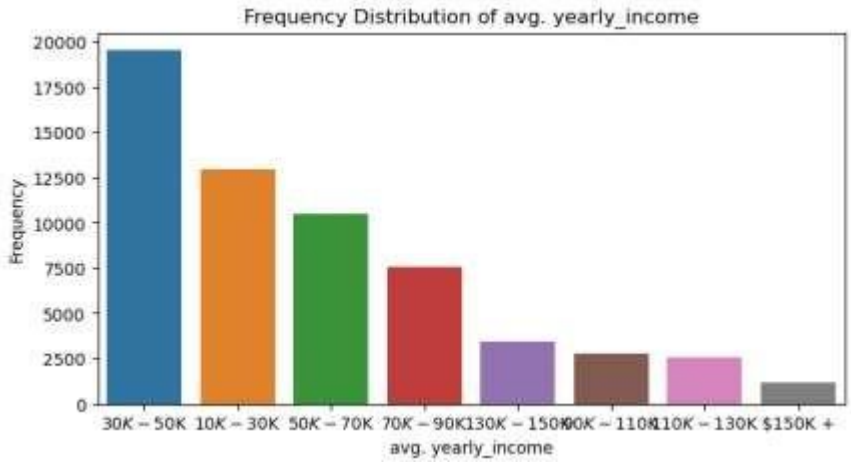
❖ **Fig 7: Sales Country (Categorical Data Type)**




The frequency distribution of the sales country is displayed above. From this data set it is inferred that the purchase is made from 3 different countries USA, Mexico and Canada. Most frequent purchases were made from USA.

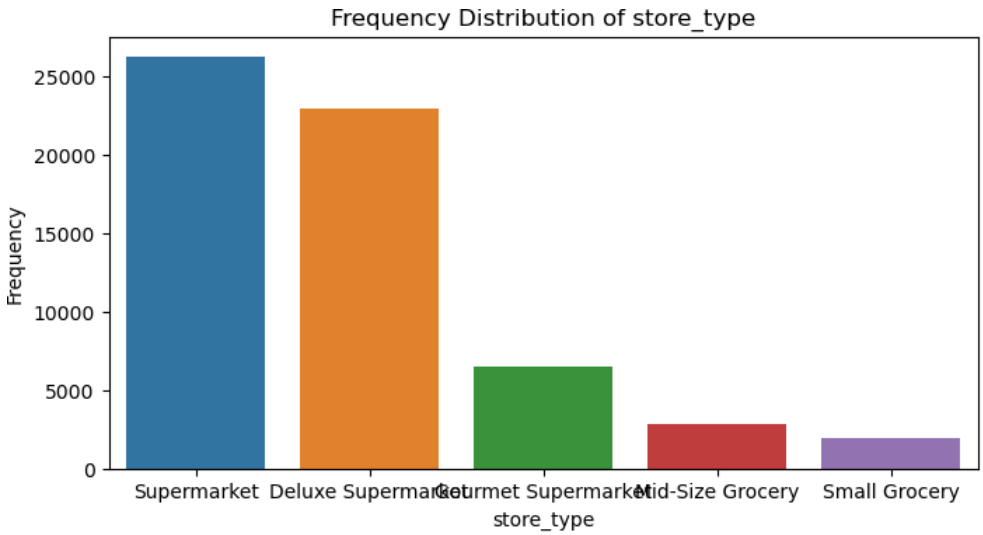
| **Sales Country** | **Frequency of Purchase** |
|---|---|
| USA | 39000 |
| Mexico | 18000 |
| Canada | 4000 |

❖ **Fig 8: Yearly Income(average) (Categorical Data Type)**




The yearly income at average is mentioned above, we can clearly observe that people having the average income between 30K to 50K have made more purchase and people having income more than 150k have made least purchase.

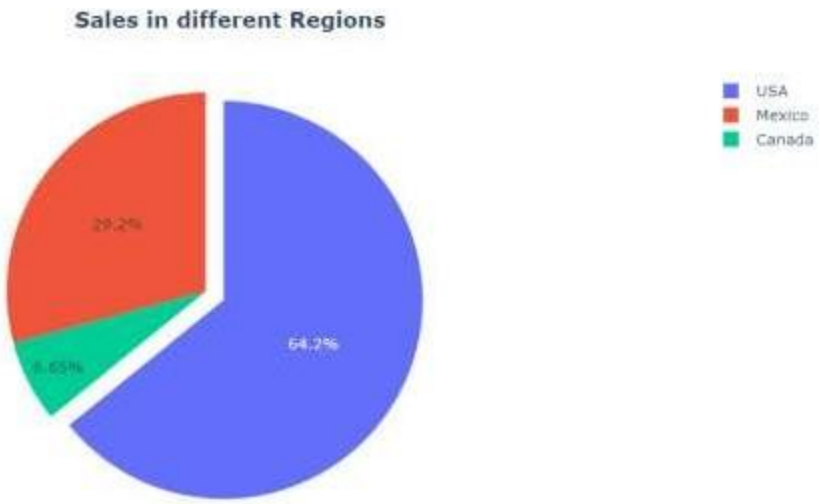
❖ **Fig 9: Store Type (Categorical Data Type)**




The stores are categorized in 5 different type Super market, Duplex Super market, Small Grocery, Mid-Size Grocery and Determent Supermarket. Most purchase were made in Super market and then Deluxe Supermarket from this we can infer that size of the shop plays a major role in attracting and retaining the customer.

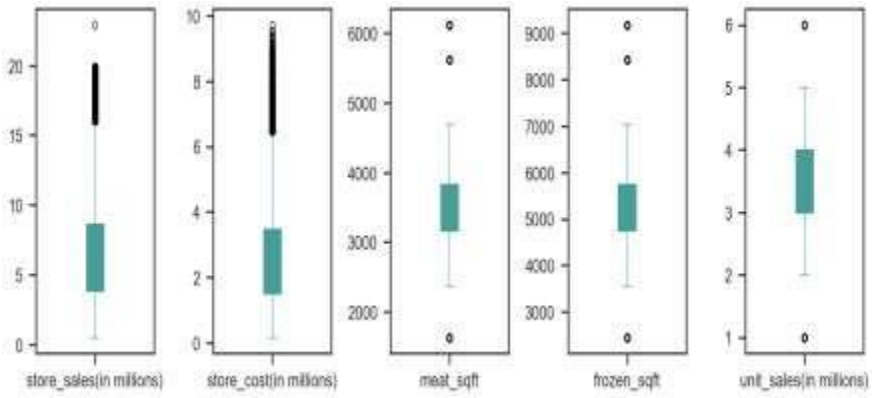
| **Shop Type** | **Number of Purchases** |
|---|---|
| Super Market | 28000 |
| Duplex Super Market | 24000 |
| Small Grocery | 3000 |
| Determent Grocery | 6000 |
| Mid-Size Grocery | 4000 |

❖ **Fig 10: Product performance based on the Region:**




as we see above that USA has the highest net salaries than the other countries cause it has the most num stores

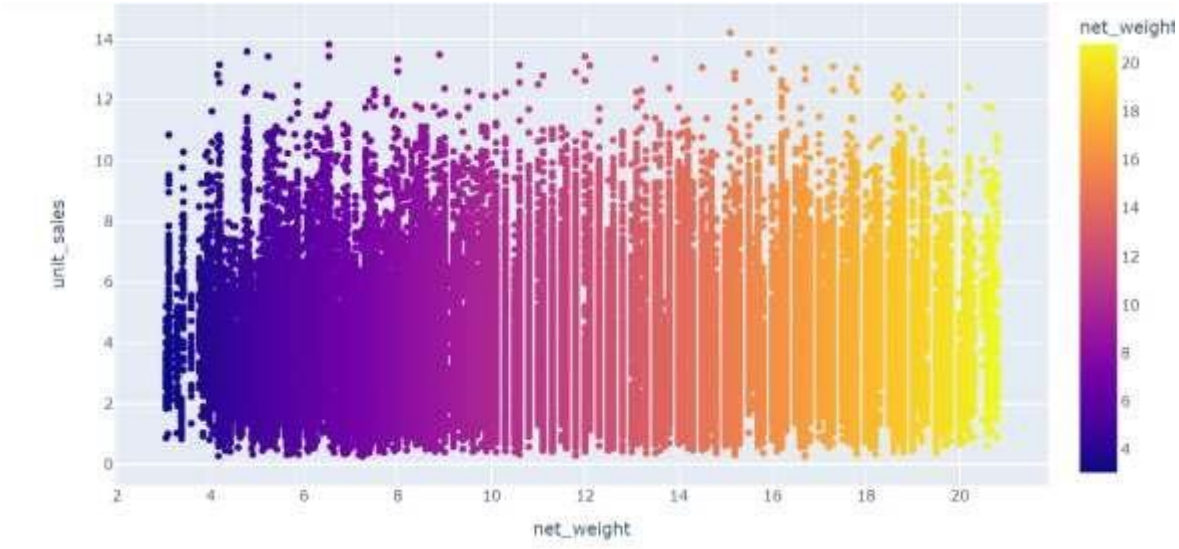
❖ **Fig 11: Box-Plot of Numerical Data Type :**




Box plot, The distribution of different data column is displayed above .From this we can infer that store sales and stores cost have more outliers and extreme value whereas meet square feet and frozen square feet and unit sales have less amount of outliers comparatively .

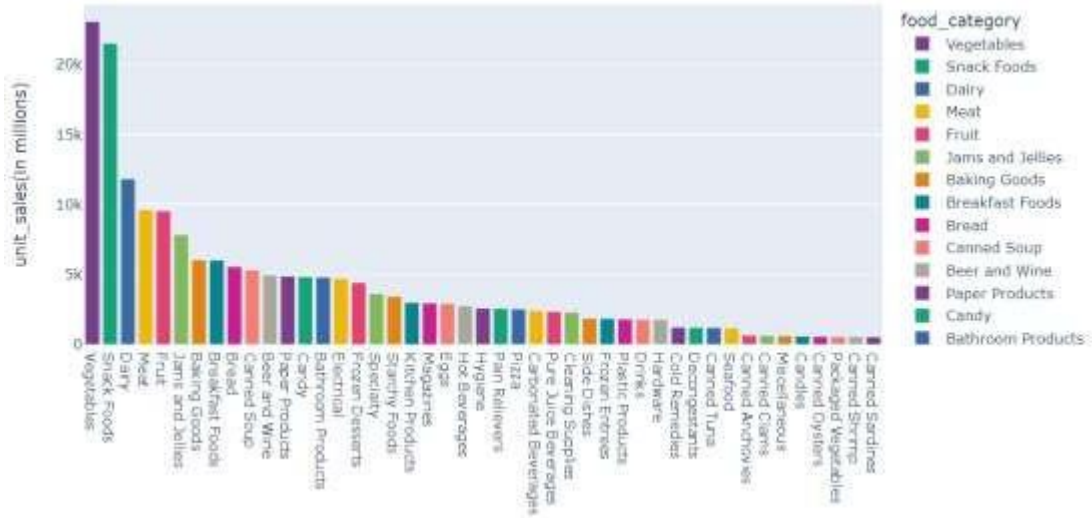
# Bivariate

❖ **Fig 12: NET weight VS Unit Sales**



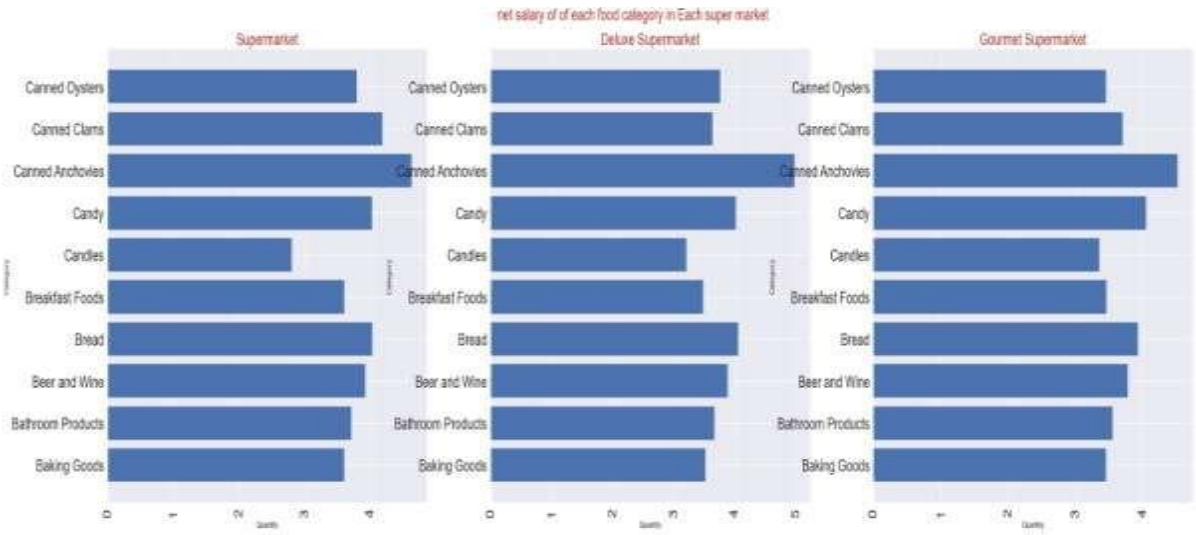

By using Group by function for each Net weight the sum of unit sales is calculated and then then relationship between them is obtained. From this we can infer that it doesnot follow any kind of pattern its neither positively or negatively correlated its just scattered .

❖ **Fig 13:Unit price for each food products**



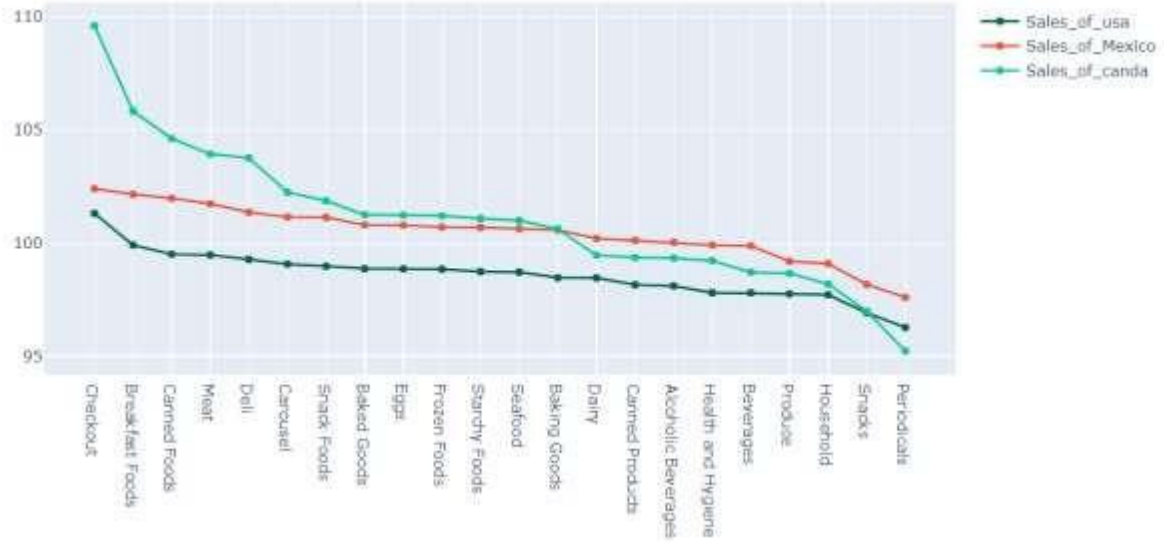

For each type of the food the total sum of unit price is displayed. From this bar chat we infer that vegetables and snack food are the most sold, it has a greater number of transactions has occurred for these two products. Hence it has achieved a proper response due to the marketing campaign.

❖ **Fig 14: Store Type Vs Food Type**



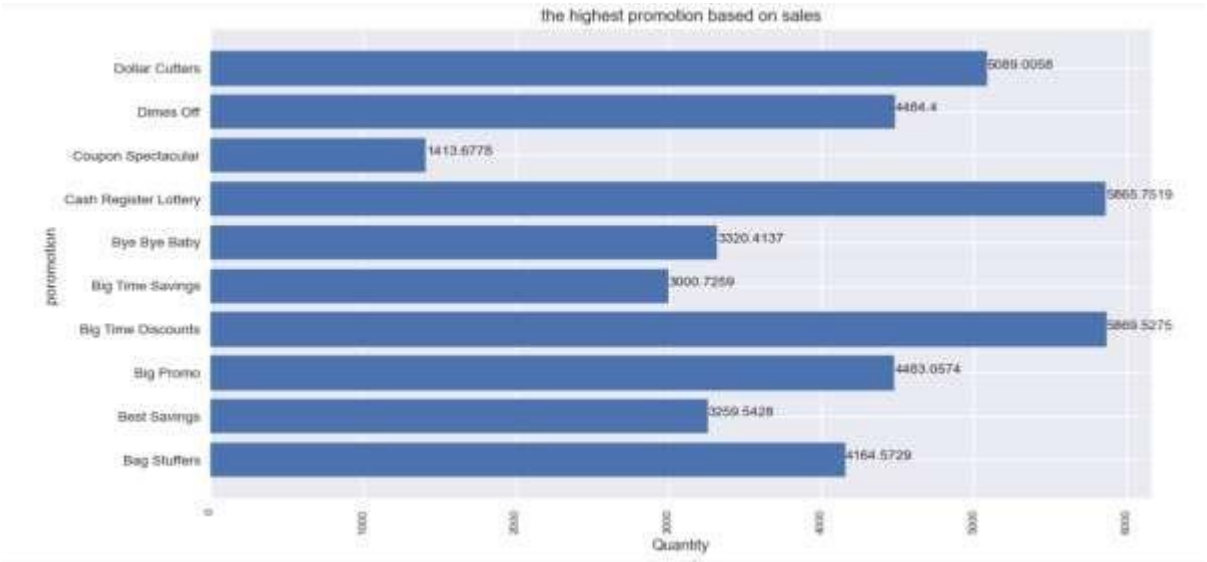

For the top 3 selling store type the top 10 selling products were displayed, so that by this analysis we focus more marketing activities on these products and so that we can see more return of investment for the amount we spend.

❖ **Fig 15 Country Sales Vs Food Type**



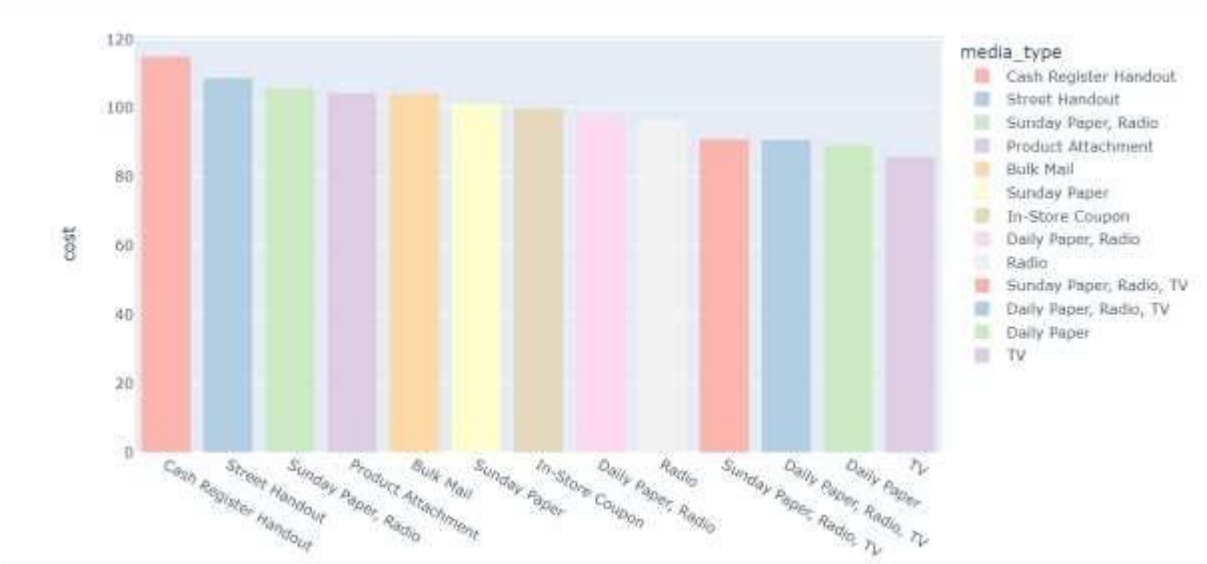

As we see from above that CANADA has the highest net sales for every food department expect from Dairy food department.

❖ **Fig 16: Highest promotion based on Sales**



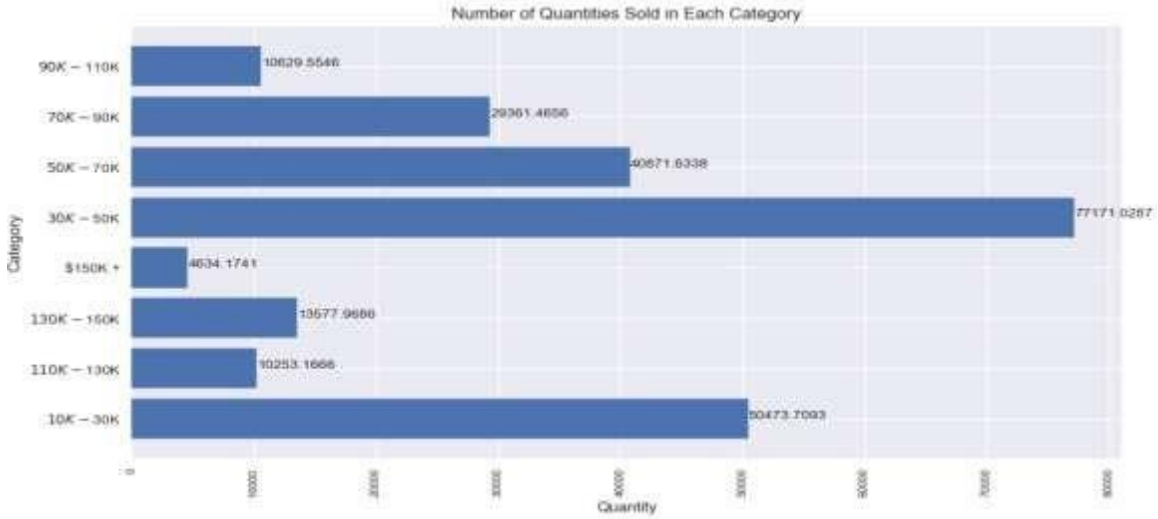

For each promotion the unit sales are calculated and displayed. It was observed that the best 4 promotions were taken among Big Time Discount , Cash register lottery , Dollar cutters and Dimes off.

❖ **Fig 17: Media type based on the cost**



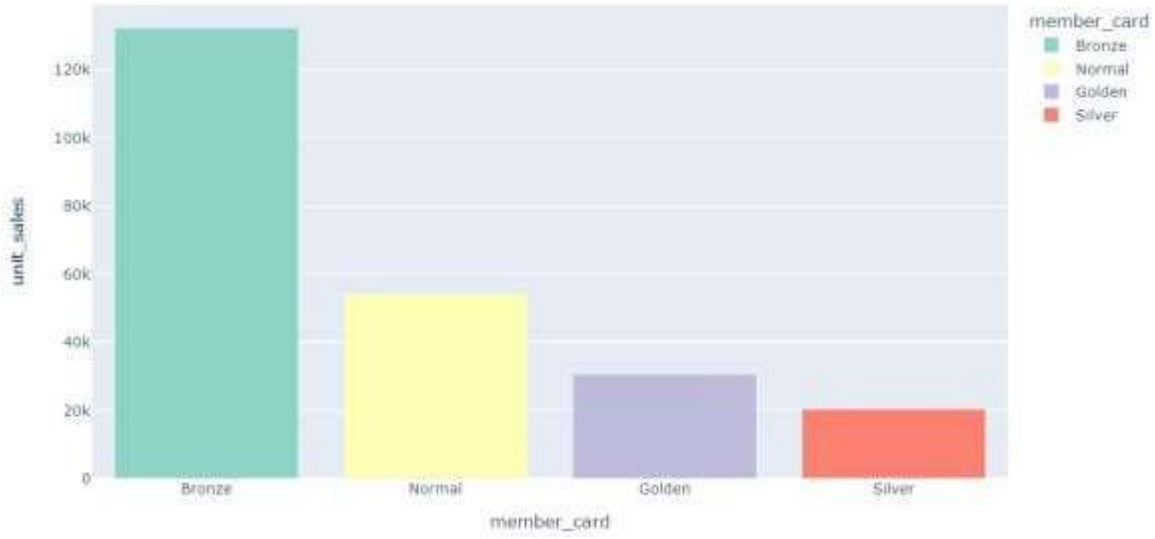

For each media type the average cost spent on this is calculated. Cash register handout is the highest cost although the daily paper, radio is the highest frequency we recommend to use TV and daily paper and radio more often

❖ **Fig 18: Yearly income Vs Unit Sales :**



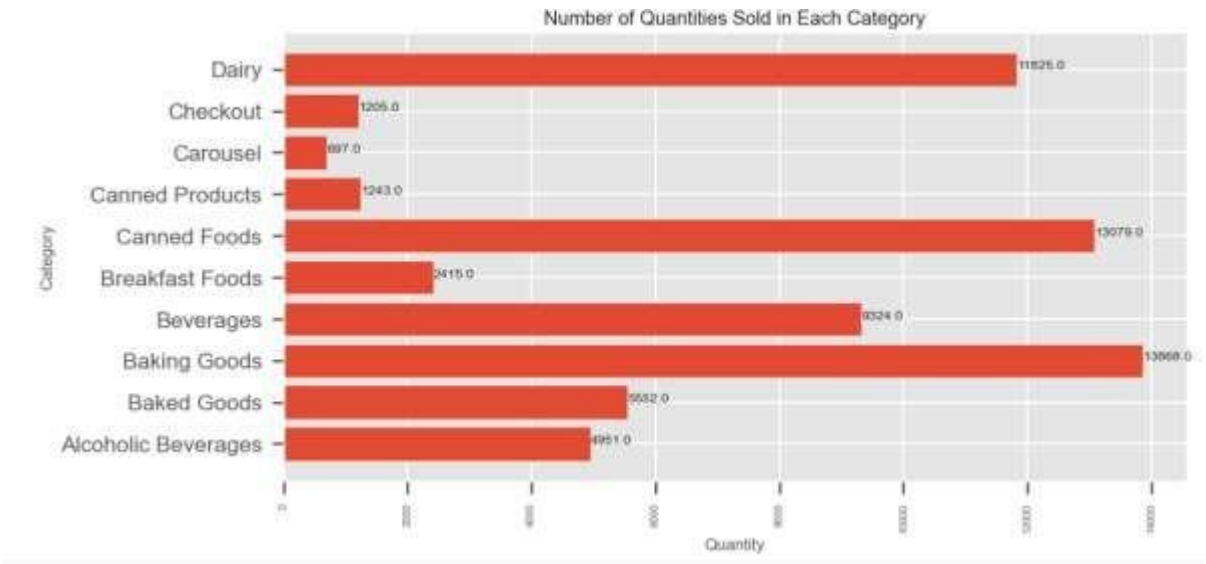

For each income category people the unit sales at average level is calculated and displayed in the bar chart .We can observe that people having salary 30K to 50K have made most of the transactions

❖ **Fig 19: Member Card vs Unit Sales :**



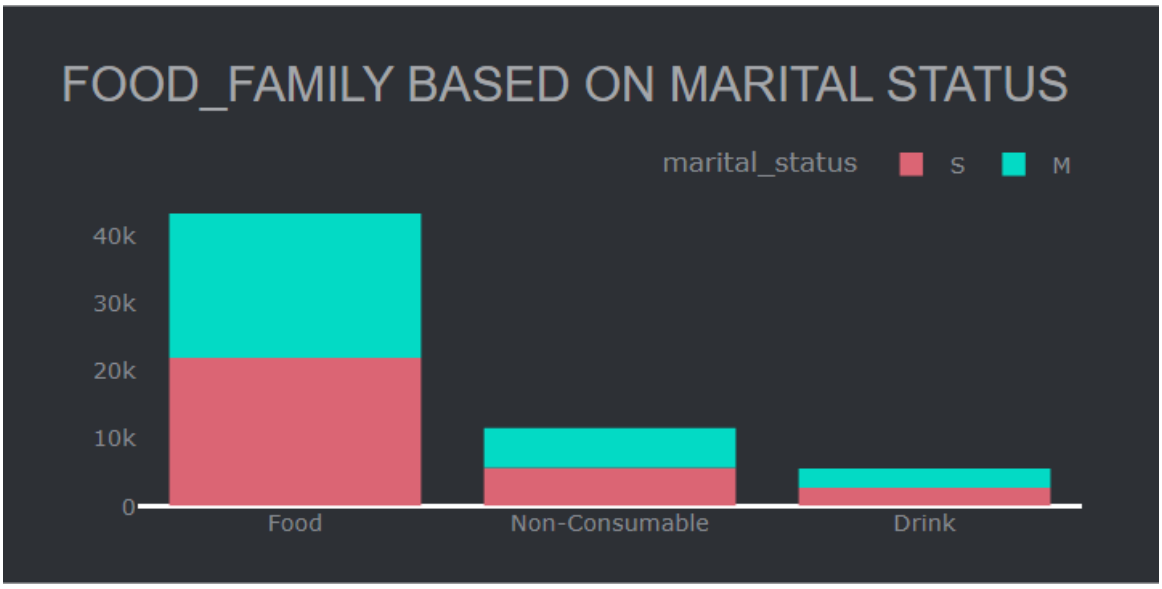

For member card category the total unit sales done is calculated and it is observed that people having bronze membership have made more transactions and silver membership have made the least.

❖ **Fig 20 : Food Department Vs Unit Sales :**



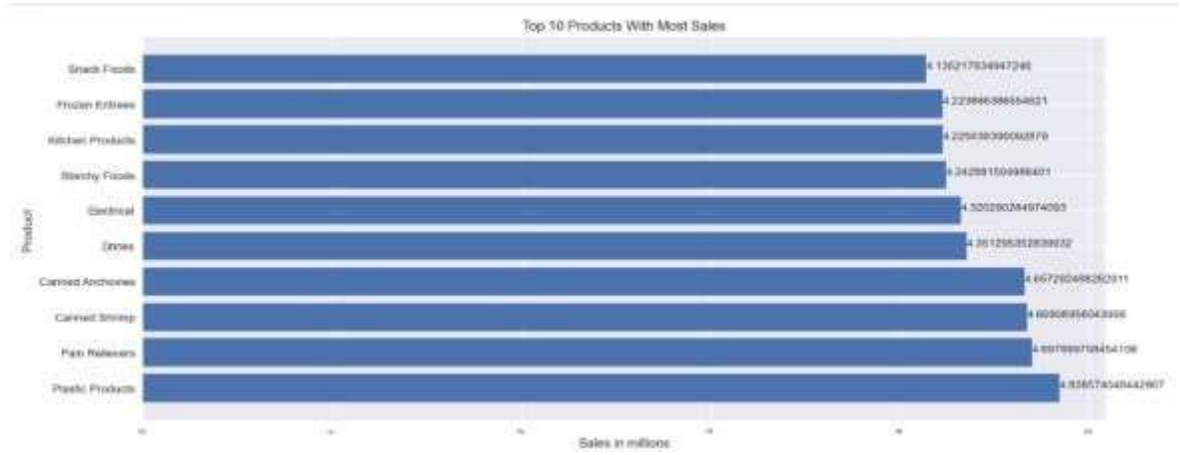

For each departments the total unit sales took was calculated and then plotted. The baking goods department achieves the highest unit sales overall the whole department

❖ **Fig 21 : Marital Status Effective on a Food family Types:**



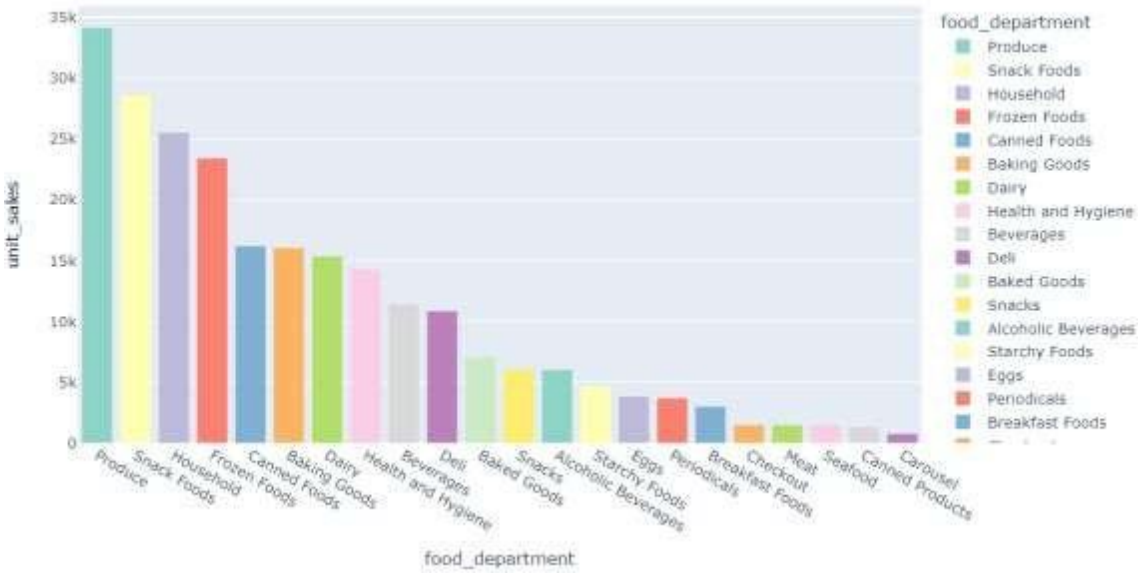

In a family, purchase of food items are more than other consumables. In that Food category, married person has made more purchase for all the three categorical food_family types.

❖ **Fig 22: Product Sales based on unit sales :**




As we see that the (pain relievers and plastic products and canned shrimp) achieves the highest net sales

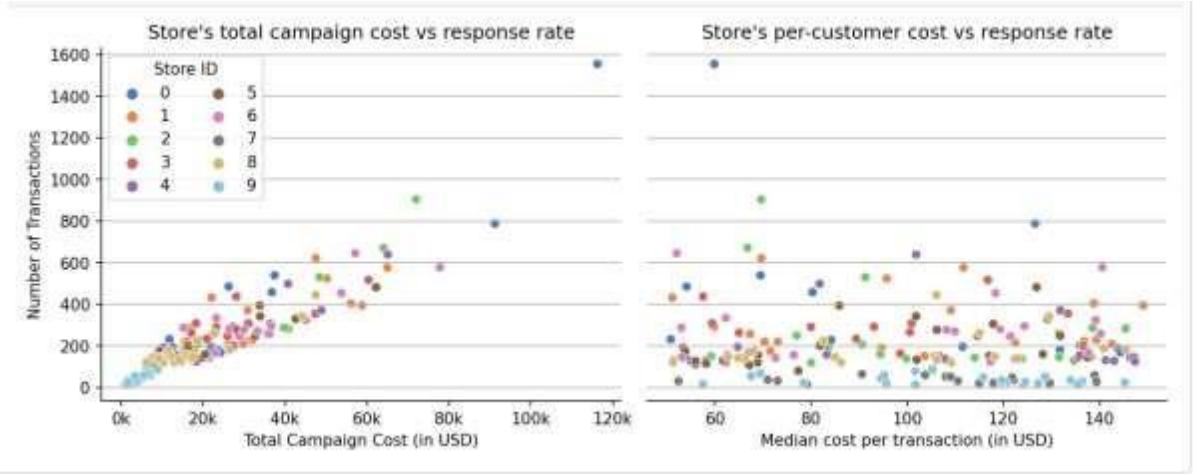
❖ **Fig 23: Food Departments based on unit sales :**




For food departments the total unit sales is calculated and top 5 food department is found
the best five department is produce, snake foods , household , frozen foods and canned foods member card with unit sales.

# Multivariate

❖ **Fig 24: Number of Transactions Vs total camping cost and media coast per transactions hue (Store ID) :**




## Inferences :

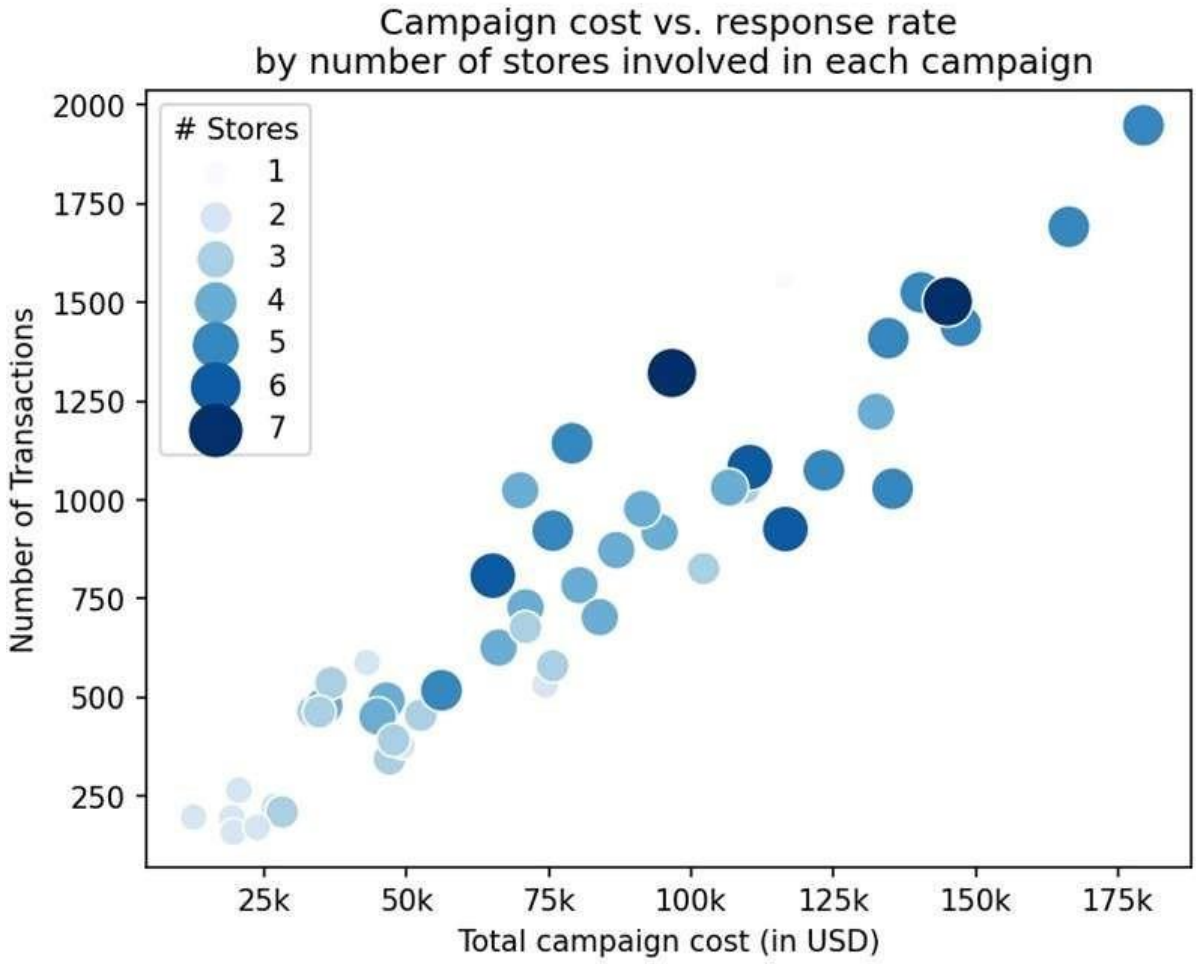
It aims to provide an insight into the efficiency and effectiveness of marketing campaigns in generating customer transactions. The x-axis, representing the "Total Campaign Cost (in USD)," ranges from 0 to 140,000 USD. In contrast, the y-axis, indicating the "Number of Transactions," spans from 0 to 1600.
The graph further seems to emphasize two crucial metrics: the "Store's total campaign cost vs. response rate" and the "Store's per-customer cost vs. response rate." These metrics are paramount for any retail establishment, as they provide a direct measure of the return on investment (ROI) for marketing campaigns. A lower per-customer cost coupled with a higher response rate would signify a more successful and cost-effective campaign.

The data on the graph seems to cluster in specific regions, which might indicate trends or patterns in the efficiency of stores' marketing campaigns. For businesses, understanding these patterns is critical. It could help in optimizing future campaigns, allocating resources more effectively, and ensuring a higher ROI.

Another metric, the "Median cost per transaction (in USD)," suggests that the graph might also be delving into the average cost incurred by the store for each transaction achieved through the campaign. This metric is crucial as it gives a direct measure of the cost-effectiveness of the campaign.

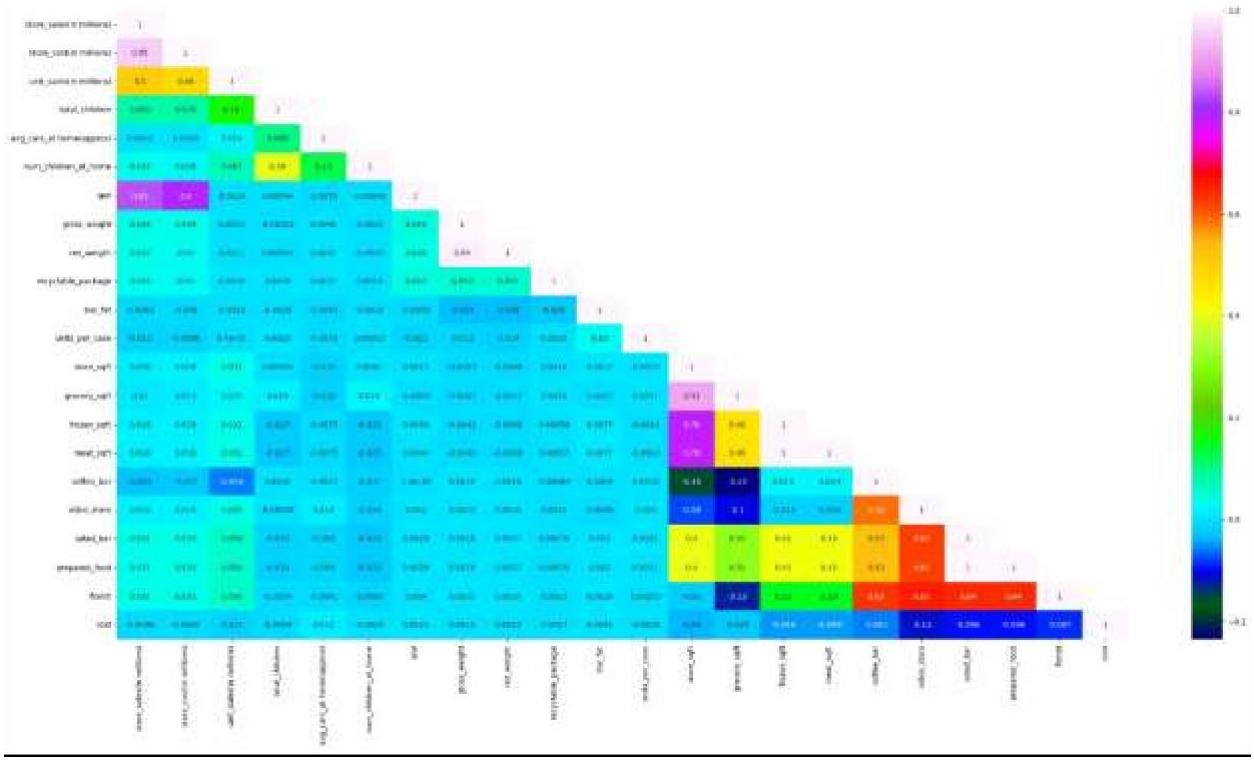
In conclusion, The significance of analysing and understanding the financial metrics associated with marketing campaigns. By the campaign cost against the number of transactions, businesses can gauge the effectiveness of their strategies and make informed decisions for future endeavours.

Fig 25 :




The image presents a graph comparing the "Number of Transactions" with the "Total Campaign Cost" for various stores, shedding light on marketing campaign efficiency. The x-axis, denoting the campaign cost, spans from 0 to 140,000 USD, while the y-axis, representing transactions, ranges from 0 to 1600. Two critical metrics stand out: the store's campaign cost versus response rate and per-customer cost versus response rate. These offer insights into the return on investment of marketing endeavors. The presence of "Median cost per transaction" further emphasizes the importance of cost-effectiveness. Overall, the graph underscores the need for businesses to evaluate their marketing strategies carefully, ensuring optimized costs and maximized transaction rates.

❖ **Fig 26 Correlation Heatmap of variables:**



# Customer Acquisition Cost Analysis

| promotion_name | clc | n_transactions | cost_per_customer | spread | levels |
|---|---|---|---|---|---|
| Bag Stuffers | 91196.56 | 788 | 126.62 | 44.00 | 2 |
| Big Time Discounts | 21855.96 | 196 | 111.51 | 0.00 | 1 |
| Big Time Savings | 9888.25 | 185 | 53.45 | 0.00 | 1 |
| Cash Register Lottery | 116211.57 | 1555 | 59.86 | 68.02 | 4 |
| Double Down Sale | 11783.28 | 232 | 50.79 | 0.00 | 1 |
| Free For All | 27183.87 | 189 | 143.83 | 0.00 | 1 |
| Green Light Days | 9741.69 | 147 | 66.27 | 0.00 | 1 |
| High Roller Savings | 19188.48 | 228 | 84.16 | 0.00 | 1 |
| I Cant Believe It Sale | 37444.33 | 539 | 69.47 | 0.00 | 1 |
| One Day Sale | 23715.00 | 180 | 131.75 | 0.00 | 1 |
| Price Savers | 36772.82 | 458 | 80.29 | 0.00 | 1 |
| Price Slashers | 26243.35 | 485 | 54.11 | 0.00 | 1 |
| Save-It Sale | 24255.84 | 168 | 144.38 | 0.00 | 1 |
| Tip Top Savings | 7027.20 | 128 | 54.90 | 0.00 | 1 |

There are 577 Products ,1798 Customer and 7 stores unique characters in past data in campaign

To find the cost required to get the customer to enter in campaign ,already they created 191 different types of promotions.

Customer Acquisition Cost (CAC) is calculated by dividing the total cost of acquiring customers by the number of customers gained.

# ABC Analysis

| | food_category | food_department | Total_Revenue | cum | cum_percent | ABC_Revenue |
|---|---|---|---|---|---|---|
| 0 | Snack Foods | Snack Foods | 47726.96 | 47726.96 | 12.074784 | A |
| 1 | Vegetables | Produce | 31777.87 | 79504.83 | 20.114494 | A |
| 2 | Dairy | Dairy | 25705.33 | 105210.16 | 26.617869 | A |
| 3 | Fruit | Produce | 17863.05 | 123073.21 | 31.137169 | A |
| 4 | Jams and Jellies | Baking Goods | 15400.78 | 138473.99 | 35.033522 | A |
| 5 | Meat | Deli | 14710.30 | 153184.29 | 38.755186 | A |
| 6 | Bread | Baked Goods | 11813.05 | 164997.34 | 41.743854 | A |
| 7 | Baking Goods | Baking Goods | 11313.59 | 176310.93 | 44.606159 | A |
| 8 | Electrical | Household | 11147.72 | 187458.65 | 47.426501 | A |
| 9 | Paper Products | Household | 10767.78 | 198226.43 | 50.150718 | B |
| 10 | Canned Soup | Canned Foods | 10419.14 | 208645.57 | 52.786731 | B |
| 11 | Vegetables | Frozen Foods | 10365.33 | 219010.90 | 55.409129 | B |

By aggregating the food category product the total revenues generated for each category is found
By sorting the total revenue and by finding the cumulated percentage obtained it is categorized Into 3 groups A,B,C

```
1 Bar_graph_Abc
```

| ABC_Revenue | Revenue | count |
|---|---|---|
| A | 187458.65 | 9 |
| B | 123696.63 | 15 |
| C | 84106.12 | 30 |

A category product has generated more revenue than B and C but count of the product is less compared to B and C. The reverse visa comes to C from this we can conclude that A generates more revenue with a smaller number of product while C with more type of products it generates very less revenue. Hence by increasing the marking campaign activities for A category product we can achieve more profit comparing to C and B is moderate to revenue and count.

- **Performance Statistical Test, Chi-Square test of Independence for Categorical vs Target Variables assuming 5% level of significance.**

| Column | p-value chi2_contingency | Result |
|---|---|---|
| food_category | 0.0067 | There is significant association between the two variables and we reject the null hypothesis |
| food_department | 0.0049 | There is significant association between the two variables and we reject the null hypothesis |
| food_family | 0.0000 | There is significant association between the two variables and we reject the null hypothesis |
| promotion_name | 0.0000 | There is significant association between the two variables and we reject the null hypothesis |
| sales_country | 0.0000 | There is significant association between the two variables and we reject the null hypothesis |
| marital_status | 0.0000 | There is significant association between the two variables and we reject the null hypothesis |
| store_type | 0.0000 | There is significant association between the two variables and we reject the null hypothesis |
| store_city | 0.4605 | There is no significant association between the two variables and we fail to reject the null hypothesis |
| media_type | 0.0000 | There is significant association between the two variables and we reject the null hypothesis |
| member_card | 0.0000 | There is significant association between the two variables and we reject the null hypothesis |
| gender | 0.6952 | There is no significant association between the two variables and we fail to reject the null hypothesis |

**Categorical and Target Variable :** To understand the relation between the categorical vs Target variable we are performing a Statistical Test namely Chi2 Contingency test for checking the significance of the categorical variable with respect to the Target Variable.

**Chi-Squared Test:** Chi-square test of independence of variables in a contingency table. This function computes the chi-square statistic and p-value for the hypothesis test of independence of the observed frequencies in the contingency table *observed*. P stands for probability here. To calculate the p-value, the chi-square test is used in statistics. The different values of p indicate the different hypothesis interpretation, are given below:

- **P >= 0.05: Hypothesis Accepted**
- **P < 0.05: Hypothesis Rejected**

**Inference:**

From the Observation Statistical Test Performed, we understand that there is no significant association between the following variables with our target variable:

1) gender
2) store_city

There is significant association between the following variables with our target variable:

1) food_category
2) food_department
3) food_family
4) promotion_name
5) sales_country
6) marital_status
7) store_type
8) media_type
9) member_card

# Feature Engineering

**Whether any transformations required**

There are many types of transformation available but we went ahead with square root transformation and power transformation. As it shrinks the data points closer and brings them to near normal distribution. So that the model's performance won't be affected by the presence of extreme values.

**Scaling the data**

For some models, we have done power transformation by default it uses the standard scaler first and transforms the data to near normal distribution, we also employed square root transformation it will shrink the data from higher magnitude to lower magnitude so that model performance won't be affected by higher magnitude values.

**Feature selection**

We will Perform this Method During Model Building as Part of Combination Top Performing Feature & Least Performing Feature

## ❖ Encoding the Categorical Variables:

The performance of a machine learning model not only depends on the model and the hyperparameters but also on how we process and feed different types of variables to the model. Since most machine learning models only accept numerical variables, pre-processing the categorical variables becomes a necessary step. Converting these categorical variables to numbers such that the model is able to understand and extract valuable information is known as Encoding. There are various encoding techniques available like Dummies, One Hot Encoder, Label Encoder, Ordinal Encoder etc., We have used Label Encoder for encoding some of our categorical variables.

We use Label encoding technique when the categorical feature is ordinal. In this case, retaining the order is important. Hence encoding should reflect the sequence. In Label encoding, each label is converted into an integer value. We will create a variable that contains the categories representing the education qualification of a person likewise.

The given below is the final processed data which we would be using in building various Classification Models like Logistic Regression, Decision Tree, Random Forest etc. We, with the help of the built models we would infer on the significance and effects of each independent variable on our target variable for predicting the patterns and rate of successful conversion to give some insightful ideas for effective marketing.

# MODELLING AND EVALUATION

**Modelling:**

The Modeling is the core of any machine learning project. This step is responsible for the results that should satisfy or help satisfied the project goals. Building a model in machine learning is creating a mathematical representation by generalizing and learning from training data. Then, the built machine learning model is applied to new data to make predictions and obtain results. Our problem statement come under the Classification thus we have decided to use
various models namely:

1. Linear Regression Model
2. Decision Tree Model
3. Random Forest Model
4. Ada Boost Technique
5. XG Boost Technique
6. Gradient Boosting Technique
7. Stacking Classifier Technique

**Value Counts of Train and Test**

**Target Variable Value Counts Train:** 48342

**Target Variable Value Counts Test:** 12086

# Model Performance & Evaluation

**Performance Score Table of Each Model**

| Model | test_size | R2_score | RMSE | MSE | MAE |
|---|---|---|---|---|---|
| **Linear Regression Train** | 0.2 | 0.312557 | 24.79 | 614.54 | 26.12 |
| **Linear Regression Test** | 0.2 | 0.302886 | 24.86 | 618.01 | 25.67 |

Our Base Model is It is overfitting Model since performance measures for training data and test data are in the same range of values. We cannot consider this model as the key assumption of absence of multi-collinearity is not satisfied.

**Parametric Model Inference:**

The performance scores for the Linear Regression model show that the model performs better on the test set than on the train set. This is not a good sign, as it means that the model is underfitting the train data. With test size 0.2 (20% of Test Data & 80% of Train Data). Need model to be better fit.

**The Linear Regression model's performance assessment reveals several key insights. Notably, the model demonstrates favorable generalization to unseen data, as it outperforms the training set on the test set, implying it isn't overfitting the training data.**

1. **Mean Absolute Error (MAE):** MAE measures the average absolute difference between the predicted values and the actual values. It quantifies how far, on average, the predictions are from the true values.

   Formula: MAE = Σ |actual - predicted| / n

   **Interpretation:** A lower MAE indicates better model performance, with smaller errors.

2. **Mean Squared Error (MSE):** MSE measures the average squared difference between the predicted values and the actual values. It penalizes larger errors more heavily than MAE.

   Formula: MSE = Σ (actual - predicted)^2 / n

   **Interpretation:** A lower MSE indicates better model performance, with smaller squared errors. However, it is sensitive to outliers.

3. **Root Mean Squared Error (RMSE):** RMSE is the square root of MSE. It provides an interpretable metric in the same units as the target variable.

   **Formula: RMSE = √(MSE)**

   **Interpretation:** A lower RMSE is preferable as it represents smaller errors.

4. **R-squared ($R^2$) or Coefficient of Determination:** R-squared measures the proportion of the variance in the target variable that is explained by the model. It ranges from 0 to 1, with higher values indicating a better fit.

   **Formula:** $R^2 = 1 - (MSE\ (model) / MSE\ (baseline))$

   **Interpretation:** A higher $R^2$ indicates a better fit of the model to the data. An $R^2$ of 1 means the model explains all the variance, while an $R^2$ of 0 means the model provides no improvement over a simple mean.

**Hyperparameter Tuning:**

By training a model with existing data, we are able to fit the model parameters. However, there is another kind of parameter, known as Hyperparameters, that cannot be directly learned from the regular training process. They are usually fixed before the actual training process begins. These parameters express important properties of the model such as its complexity or how fast it should learn.

Models can have many hyperparameters and finding the best combination of parameters can be treated as a search problem. The best strategies for Hyperparameter tuning are:

1. **GridSearchCV**

   In GridSearchCV approach, the machine learning model is evaluated for a range of hyperparameter values. This approach is called GridSearchCV, because it searches for the best set of hyperparameters from a grid of hyperparameters values.

**Non-Parametric Model:**

1. **Decision Tree Regressor:**
   - A Decision Tree Regressor is a supervised machine learning algorithm used for regression tasks. It creates a tree-like structure where each internal node represents a feature, and each leaf node represents a predicted numerical value.
   - The decision tree splits the data based on feature values to minimize the mean squared error (MSE) or other regression loss functions.
   - Decision Trees can capture non-linear relationships in data and are interpretable.

2. **Random Forest Regressor:**
   - A Random Forest Regressor is an ensemble learning method that combines multiple decision tree regressors to make more accurate predictions.
   - It fits a collection of decision tree regressors on different subsets of the dataset and averages their predictions to reduce overfitting and improve predictive accuracy.

3. **Gradient Boosting Regressor:**
   - Similar to the Gradient Boosting Classifier, the Gradient Boosting Regressor builds an additive model in a forward stage-wise fashion for regression tasks.
   - It fits a sequence of regression trees to the negative gradient of the loss function, optimizing for regression loss functions like mean squared error (MSE).

4. **XGBoost Regressor:**

    - XGBoost can also be used for regression tasks. It shares many features with the XGBoost Classifier but is applied to predict continuous numerical values.
    - XGBoost includes enhancements like parallel computing, cache optimization, and regularization to improve regression accuracy and prevent overfitting.

5. **AdaBoost Regressor:**

    AdaBoost (Adaptive Boosting) is an ensemble learning technique that can be applied to regression problems as well. The AdaBoost Regressor works similarly to the AdaBoost Classifier but focuses on improving regression performance. It combines multiple weak regression models (usually decision trees with limited depth) to create a strong ensemble model.

6. **Stack Classifier (Stacking):**

    Stacking is an ensemble learning technique used for classification problems, but it can also be adapted to regression tasks (Stacking Regressor). Stacking combines predictions from multiple diverse models to improve overall predictive performance.

# Base Model – Hyperparameters Tunned

| S.NO | Model Name | Train Data | | | | Test Data | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | R-Square | RMSE | MSE | MAE | R-Square | RMSE | MSE | MAE |
| 1 | Base Model (OLS) | 0.318719 | 24.7957 | 614.8267 | 24.4543 | 0.307996 | 24.86429 | 618.2329 | 24.8643 |
| 2 | Linear Regression (Base Model) | 0.318719 | 24.7957 | 614.8267 | 24.9874 | 0.307996 | 24.86429 | 618.2329 | 24.8643 |
| 3 | OLS model with power Transformer | 0.31629 | 24.83981 | 617.0162 | 24.7883 | -1064017 | 9749817.37 | 9.51E+13 | 9,751,922.89 |
| 4 | Linear regression with power Transformer | 0.31629 | 24.83981 | 617.0162 | 24.4783 | -1064017 | 9749817.37 | 9.51E+13 | 9,751,922.89 |
| 5 | Linear Regression with VIF removal | 0.020975 | 29.72421 | 883.528 | 29.9563 | 0.02292 | 29.54506 | 872.9106 | 29.5451 |
| 6 | Linear Regression with SFS best (Forward ) | 0.31866 | 24.796760 | 614.8793 | 24.2123 | 0.31866 | 24.86328 | 618.1827 | 24.8633 |
| 7 | Linear Regression with SFS best (Backward) | 0.31864 | 24.797050 | 614.8937 | 24.4324 | 0.30798 | 24.79705 | 614.8937 | 24.7971 |
| 8 | RFE(Recursive Feature Elimination) | 0.01445 | 29.82302 | 889.4125 | 29.9563 | 0.01754 | 29.62633 | 877.7194 | 29.6263 |
| 9 | Decision Tree Regressor | 1.0 | 2.72766 | 7.440129 | 2.761 | 0.9987 | 2.660831 | 7.080022 | 2.6608 |
| 10 | Decision Tress Regressor (With best Param) | 0.99083 | 2.87579 | 8.270168 | 2.898 | 0.99091 | 2.84901 | 8.116858 | 2.8490 |
| 11 | Decision Tree with SFS | 1.0 | 2.6601441 | 7.076367 | 2.631 | 0.99248 | 2.660144 | 7.076366 | 2.6601 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | (forward Best) | | | | | | | | |
| 12 | Decision tree with SFS (Backward Best) | 1.0 | 2.16934 | 4.706036 | 2.621 | 0.9897 | 2.66014 | 7.076345 | 2.6601 |
| 13 | Decision Tree (grid search CV) | 1.0 | 2.72766 | 7.440129 | 2.6312 | 0.99207 | 2.66567 | 7.105797 | 2.6657 |
| 14 | Random Forest Regressor | 0.9992221 | 0.8378451 | 0.701984 | 2.712 | 0.994464 | 2.2237765 | 4.945182 | 2.2238 |
| 15 | Random Forest Regressor with SFS (forward) | 0.99923 | 0.83337490 | 0.694514 | 2.343 | 0.99923 | 2.1047559 | 4.429997 | 2.1048 |
| 16 | Random Forest Regressor with SFS (Backword) | 0.999212 | 0.843020 | 0.710683 | 2.623 | 0.994573 | 2.2018013 | 4.847929 | 29.224 |
| 17 | Random Forest Regressor With RFE | 0.053296 | 29.229444 | 854.3604 | 28.778 | 0.051980 | 29.229444 | 854.3604 | 28.778 |
| 18 | Random Forest Regressor with Grid Search CV | 0.999222 | 2.22377 | 4.945153 | 2.2018 | 0.98784 | 2.37787 | 0.38782 | 0.6228 |
| 19 | Random Forest Regressor with best feature (using GS CV) | 0.9908 | 2.4567 | 6.035375 | 2.6571 | 0.9808 | 2.5765 | 0.4989 | 0.7063 |
| 20 | Ada boost Regressor | 0.381915 | 23.617689 | 557.7952 | 23.6177 | 0.380766 | 23.617689 | 557.7952 | 23.6177 |
| 21 | Gradient Boost Regressor | 0.381915 | 23.61768 | 557.7948 | 23.617 | 0.38076 | 23.61768 | 557.7952 | 23.617 |
| 22 | XGB boost Regressor | 0.9929 | 2.523597 | 6.368542 | 0.3212 | 0.990718 | 2.879620 | 5.891 | 2.4271 |
| 23 | XGB with Grid Search CV | 0.955 | 6.320 | 39.943 | 0.3437 | 0.9534 | 6.23 | 39.945 | 0.3545 |

## Feature Importance for the best Model After Hyperparameters Tunned:

**Model:** XGBoost Classifier With GridSearchCV

**Parameter Used**:

**GridSearchCV**

| | |
|---|---|
| **Max Depth** | 4, 5, 6 |
| **n_estimator** | 500, 600, 700 |
| **Learning_rate** | 0.01, 0.015 |

**Best Parameters**

| | |
|---|---|
| **Criterion** | Entropy |
| **Max Depth** | 6 |
| **n_estimator** | 700 |
| **Learning_rate** | 0.015 |

**Model Evaluation Chart Based R2 Score**

| **Model .No** | **Model** | | |
|---|---|---|---|
| | | Training dataset | Test dataset |
| **1** | Decision Tree Regressor | 1 | 0.960316 |
| **2** | Random Forest Regressor | 0.980876 | 0.981542 |
| **3** | Gradient Boost | 0.980876 | 0.981542 |
| **4** | Ada Boost | 1 | 0.961393 |
| **5** | XG Boost | 0.98512 | 0.98412 |

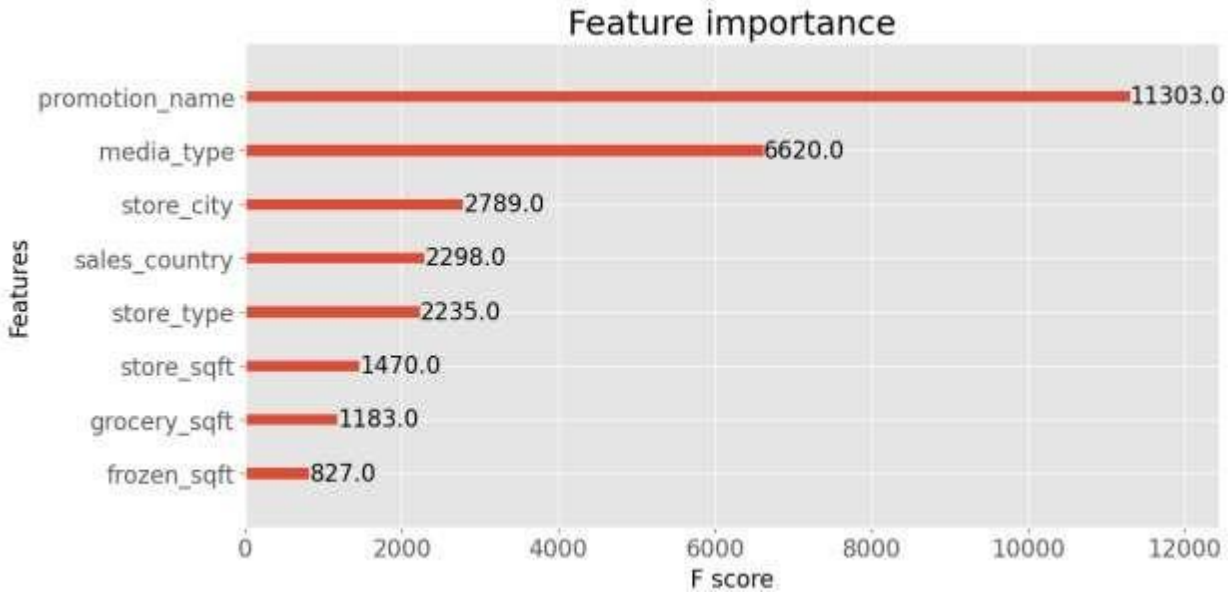
**Table.2.** R2 Score for the models Before Hyperparameter tuning

| **Model .No** | **Model** | R2 Score | |
|---|---|---|---|
| | | Training dataset | Test dataset |
| **1** | **XG Boost Classifier Train Best Param** | 0.9929 | 0.9907 |

**Table.3.** R2 Score for the models after Hyperparameter tuning

## Feature Importance Chart:

**Best Model Top 8 Feature Importance Bar Chart**




**Conclusion (Parametric Model):**
Since the key assumption of the Linear Regression model, namely absence of multi-collinearity is not satisfied, we are not considering this for our project.

**Conclusion (Non-Parametric Model):**

The XG Boost Regressor with the best parameters, stands out as the best model for further development. It exhibits exceptional r2 score of approximately 98.45% on the test data, meaning it effectively identifies.

## Business Interpretation

1. **Cost-Efficiency:** With such a high recall value, the model is excellent at identifying the right customers who are likely to respond positively to our marketing campaign. This means that our marketing resources and budget are being used efficiently, as we are not wasting resources on customers who are less likely to convert.
2. **Lower Customer Acquisition Cost:** High recall implies that we are capturing a significant

portion of potential customers without missing out on many of them. As a result, our customer acquisition cost is likely to be lower, as we are not overspending on marketing efforts that don't yield positive results.

3. **Improved ROI:** The model's accuracy in identifying potential customers can lead to a higher return on investment (ROI) for our marketing campaign. By targeting the right audience, we can expect a better conversion rate and revenue generation, thus maximizing the effectiveness of our marketing spend.
4. **Optimized Resource Allocation:** With the XGBoost Regressor's capabilities, we can make data-driven decisions about where to allocate our marketing resources. This may involve focusing more on specific channels, demographics, or strategies that the model has identified as particularly promising.
5. **Continuous Improvement:** While the XGBoost Regressor is performing exceptionally well, it's important to note that there is always room for improvement. Regularly updating and fine-tuning the model with new data and insights can further enhance its performance and help us maintain a competitive edge in customer acquisition.