# Introduction

In this project, I was tasked to examine a data set based on customer transactions, browsing behaviours, feedback on purchases, and responses to previous marketing campaign. The outcome of this report will be to provide an in depth analysis of the dataset and show a range of techniques and strategies that could be used during a targeted marketing campaign. These campaigns are personalised while being based on the outcome of the analysis.

All the statistical and coding work in this report was completed in Jupyter Notebook.I utilised a variety of analytical skills in the its completion. These include, but are not limited to EDA, descriptive statistics, probabilities and probability distribution.

# Overview

I began by performing some EDA (exploratory data analysis.) tasks. The data dictionary stated there were multiple columns that had null values but after executing the *".unique()"* function on these columns I found there to be no missing data. However, there were zeros present. As these are still numerical values and not null, I left them the way they were. The could be said for "-" values which the data dictionary said there a few but none were located upon examination.

The main EDA commands I used were as follows:
.head() – to show the first 10 observations and have a what features are present
.shape() – show in numerical format how many observations and features
.info() – Check the data types and confirm no missing values
.isnull().sum() – show the number of missing values per feature
.describe() – make a table from the mean, median, standard deviation etc…
.unique() – show the unique values in the selected feature

After completing EDA, I conducted the first part of the testing with an eye towards improving customer retention based around the variables of customer feedback scores, purchase frequency and campaign response rate.

## Part 1

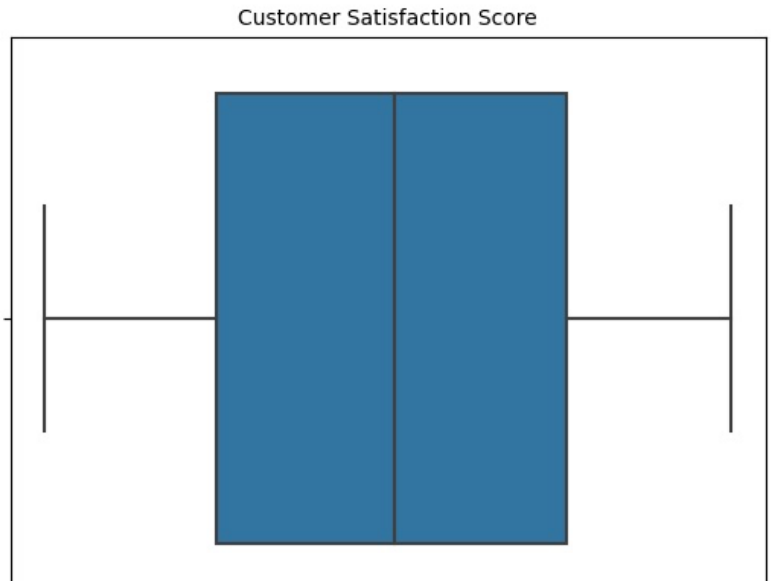


*Figure 1 - Box Plot of Customer Satisfaction Score*

Customer Satisfaction Score

**For customer feedback scores, the features I chose were:**

### Customer Satisfaction Score

The data is continuous. The most typical score is 50.21 while median is 51. The majority of the people used for the campaign are scoring only half on satisfaction. I used the customer satisfaction as this can be used to target people that aren't satisfied which could help customer retention by fixing any issues that may cause further problems for potential customers in the future. The satisfaction score would be a useful metric to build a

marketing campaign.  Establishing the cause of customer unrest would indicate an area that could be targeted. As you can see from the boxplot, the data is evenly distributed and has no outliers.

## Product Rating

The data is continuous but having consulted the data dictionary, I surmised that the "Product rating" category had to be scaled from values ranging from 0-100 to values ranging from 1-5. I used the Min-Max scaler from Sklearn for this. After scaling I found that 45.5% of people in the product rating category gave a rating of between 3 and 5.

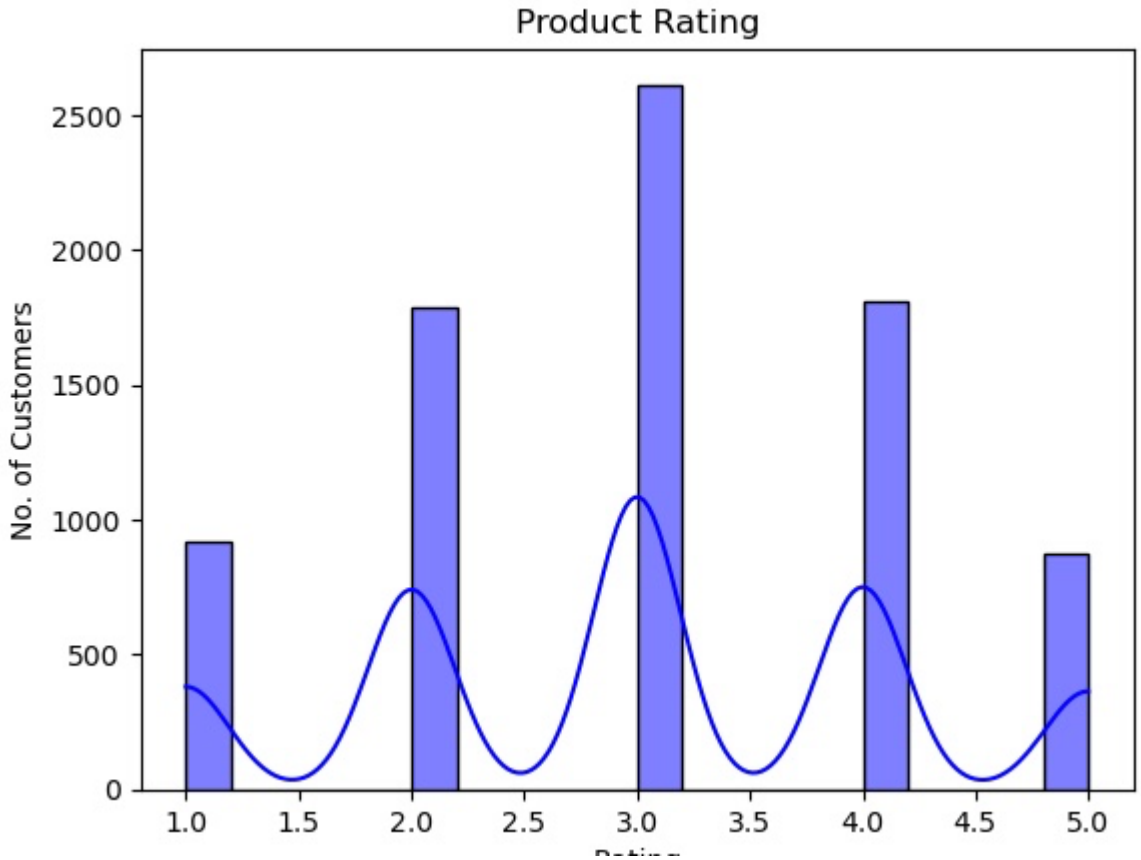


*Figure 2 - Histogram of product rating after scaling*

Understanding how customers are rating the products could create scope for a marketing campaign. For example, any customer that rated a product highly could be targeted with similar products. On the contrary, any ratings below a certain score would indicate what not to send a particular customer in terms of marketing.

**For purchase frequency, the features I chose were:**

## Transaction Amount

The data was continuous. The minimum amount spent is $0.01 and the max is $99.97. The average is $50. The transaction amount gauges loyalty and is a great metric to utilise for marketing. Customers spending over a certain amount should be rewarded with some kind of marketing bonus. 14.8% of customers have spent between $75 and $100.

## Account Age

The data is continuous. The mean length of account is 49 days with the first and third quartile falling at 25 and 75 days respectively.  The most frequent account age is 70 days. Similar to transaction amount, this is a good

variable to determine customer loyalty which can indicate various marketing campaigns such as rewards for customers with accounts over a certain amount of days.

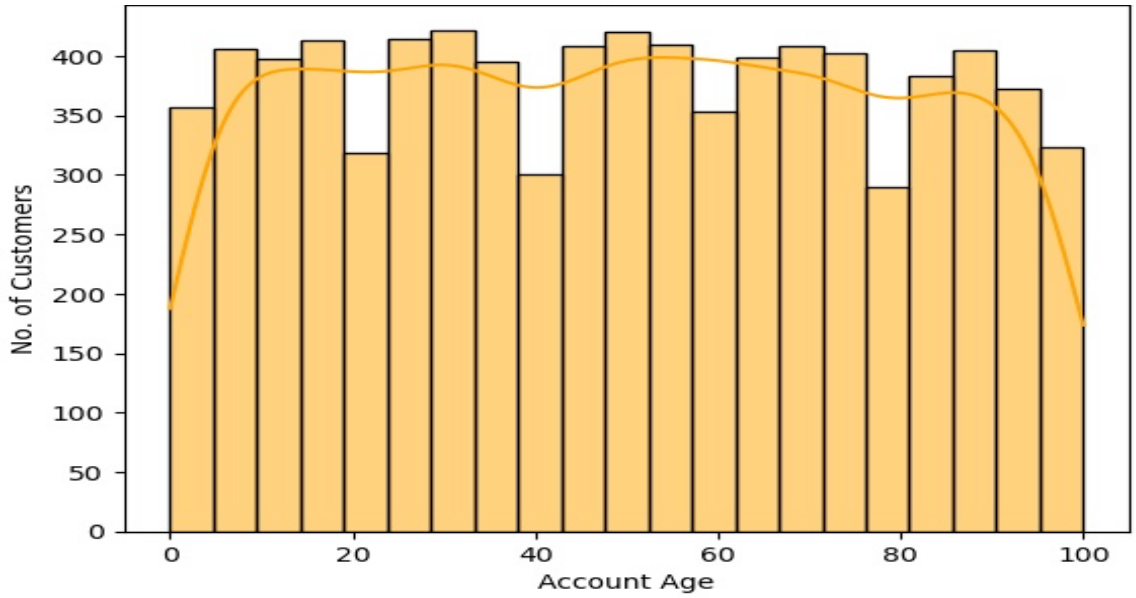


*Figure 3 - Histogram containing account age data*

To get a rough idea of purchase frequency, the transaction amount could be divided by the account age. This would give an average of what each customer is spending per day. Customers could be targeted if their average spend per day is above a certain amount.

**For campaign response rate, the features I chose were:**

## Marketing Email Click Rate

The data is a continuous percentage value. The mean is 49.41 and the standard deviation is 29.72. 14.7% of customers have a click rate of between 75% to 100%. The variation coefficient is 58% which means the values are not well dispersed around the mean.

## Newsletter Open Rate

The data is a continuous percentage value. The min value is 0% and the max is 100%. 48.3% of customers have opened between 40% - 80% of the newsletters.

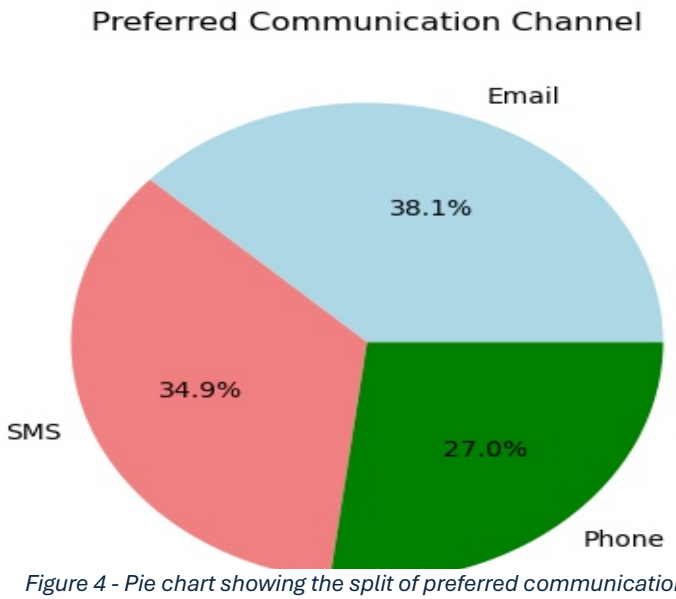


*Figure 4 - Pie chart showing the split of preferred communication*

Preferred Communication Channel
As this data is categorical there is no numerical mean value. However, I calculated the mode to be email which means the majority of people preferred to be contacted by email.

## Part 3

The most important variable to examine in terms of customers responding to future marketing initiatives would be the *"Interest in new product"* column. I concluded that probability of a customer from the dataset that would be interested in new products is 50.6%. That's just over 4049 potential customers to target with new products.

The second variable I chose here was *"Membership Status*". By calculating what percentage of people are/aren't a member I can get an idea of how customers will respond. Of the 8000 customers in this dataset, there were 7895 members and 105 non-members. This probability of there being a customer from the dataset that is a member is 98.6%.

Finally, I chose the Product *"Category Preference"* as my third variable. Knowing what kind of products customers like will be a huge benefit to marketing team. The numerical breakdown is as follows:

**Gold = 3162**
**Platinum = 2783**
**Basic = 2055**

This equates to a probability of:

**Gold = 39.5%**
**Platinum = 34.7%**
**Basic = 25.6%**

Marketing campaigns can be built around the above data. The marketing team know that on items in the gold category, the probability of a customer purchase is 39.7%. From items in the platinum category, the probability of a customer purchase is 34.7%. Finally, from items in the basic category, the probability of a customer purchase is 25.6%.

## Part 4

To criteria to determine and calculate binomial distribution is as follows:

X = Number of Elements from a particular characteristic or element within a certain limit.

The customer segment I chose to analyse was segment A. For a binomial distribution, I need three parameters:

**K = number we are looking for**
**N = sample size**
**P = probability**

In this case, I calculated that the breakdown of customers per segment is:

**A - 1998**
**B – 2028**
**C - 1978**
**D – 1996**

I want to calculate what is the percentage chance of someone from segment A having an interest in new products. I know that the probability, of the entire sample having interest in new products is 50.6% or 0.506125 so this my p value.

I want to take 5% of my sample size so that equates to 5% of 1998 which comes to 99.9 customers which I round up to 100. This is my n value.

The k value is a random number within the sample so I have chosen 40,50,60 and 70. The results can be seen in the table below:

| K | N | P | Result |
|---|---|---|---|
| Less than 40 | 100 | 0.506125 | 0.021349874311732 |
| Greater than 40 | 100 | 0.506125 | 0.978650125688267 |
| Exactly 40 | 100 | 0.506125 | 0.008423990910993 |
| | | | |
| Less than 50 | 100 | 0.506125 | 0.49086794752263 |
| Greater than 50 | 100 | 0.506125 | 0.50913205247736 |
| Exactly 50 | 100 | 0.506125 | 0.078994259635079 |
| | | | |
| Less than 60 | 100 | 0.506125 | 0.976356298840233 |
| Greater than 60 | 100 | 0.506125 | 0.999999997260681 |
| Between 40 - 60 | 100 | 0.506125 | 0.955006424528501 |
| | | | |
| Less than 70 | 100 | 0.506125 | 6.110255306545483 e-06 |
| Greater than 70 | 100 | 0.506125 | 0.999993889744693 |
| Between 30 - 70 | 100 | 0.506125 | 6.110255306545483 e-06 |

*Figure 5 - Table containing binomial distribution probibilities*

As you can see from the table the probability table, by choosing around 60 customers from segment A, there is a very strong probability that they'll have an interest in new products. This is an ideal way to structure a marketing campaign towards these 60 or so customers as I know they are almost guaranteed to be interested in the potential new products that are on offer.

## Part 5

The main variable I have chosen is Transaction Amount as it is normally distributed. To build a marketing campaign around this data, I wanted to target customers that have spent less than $50. To apply normal distribution I need three things: mean (mu), standard deviation (sigma) and the number I want to test up to. For this case, the *"Transaction Amount"* mean and standard deviation can be seen below.

**mu = 50.281659**
**sigma = 27.280464**

The probability of customers spending $50 or less from our sample is 0.49588116655619724 or 49.6%. This can be seen in Figure 6.

This information can be used to target valuable customers. Customers that spend more money are considered to be more valuable. However, in the interest of customer retention, I believe that it would be easier to retain more people that spend less money as oppose to less people that spend more money. Customers that spend less money but on a more regular basis are the focus of this marketing campaign and should be targeted with

products that are under a certain amount. They should also be rewarded with discounts based on the amount of smaller items purchased which would lead to customer retention.

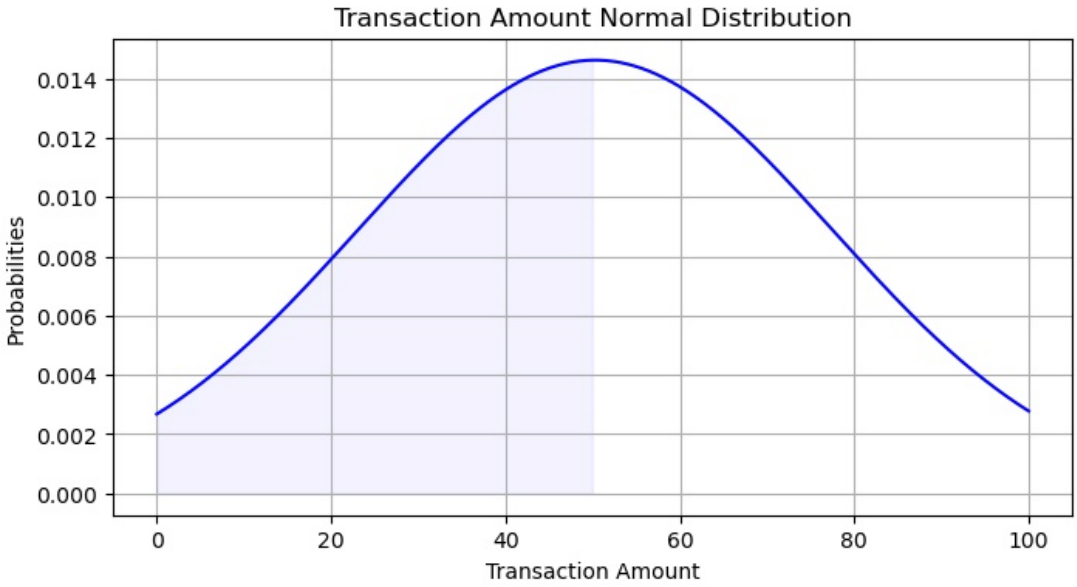


*Figure 6 - Normal distribution graph of transaction amount*