
**BT4211 Data-Driven Marketing**

**Group Project Report**

AY 2021/2022 Semester 2

**Prepared by: Group 10**

Carine Tan Kailin A0189558E

Goh Jia Yi A0185610J

Koh Gladys A0188320H

Koh Min A0188301J

Loh Zi Ying A0188859Y

**Table of Contents**

# 1 Introduction

## 1.1 Background

Flipkart is currently the largest e-commerce retailer in India, with over 160 million monthly active users and 150 million products (Sahni, 2019). As of 2020, Flipkart has a market share of 31.9% within the country but is followed closely by Amazon whose market share stands at 31.2% (Dayalani, 2021).

While e-commerce is a booming industry in India, both companies are reporting losses despite their increased sales due to India's market preference for discount-led purchasing. In an attempt to capture the market, Flipkart experienced a high cash burn in its advertising to attract new customers and retain existing customers, which led to a rise in customer acquisition and retention costs (Khatri, 2019; Poojary & Ranjan, 2019). In 2016, it is reported that Flipkart's monthly cash burn was about 2.6 billion Indian Rupee (INR) while Amazon was losing about 6 billion INR per month, more than double of Flipkart (Shrivastava et al., 2017). However, Amazon has been dumping billions of dollars into the Indian market to create the same monopoly it has in the US, making it an uneven playing field for e-commerce companies.

In order to retain its position as India's largest e-commerce retailer and stay ahead of competition, it is crucial for Flipkart to plan and utilise its marketing budget more efficiently. This can be achieved by designing customised marketing programs to target specific customer segments based on their identified underlying needs. With more data-driven decisions, it can help to reduce Flipkart's overall customer acquisition and retention costs, which translates to an increase in profit.

## 1.2 Objective

By analysing past customer transactional data, the purchase history can provide managerial insights on the customers. Customer segmentation helps to group customers with similar characteristics as well as Customer Lifetime Value (CLV), which provides an indication of how much value customers bring to the company in their entire stream of lifetime purchases.

With a better understanding of the customers and their value contribution, marketing strategies can then be designed for attractive segments that have been identified while ensuring that marketing costs are allocated efficiently to reap the highest returns.

## 1.3 Research Question

How should Flipkart segment its customers and design targeted marketing strategies?

# 2 Data Gathering and Cleaning

The Flipkart dataset[1] was obtained from Kaggle, where it contained the Purchase data and Psychometric and Demographic data of 304 customers between September 2011 and December 2021.

[1] https://www.kaggle.com/chandanmalla/marketinganalyticsdata

The Purchase data was collected through invoices shared by customers, which can be retrieved either through email or downloaded from Flipkart's website. Customers' purchase information was extracted automatically from the invoice files using Natural Language Processing (NLP) techniques. Each row represents a different item purchased within an invoice. Meanwhile, the Psychometric and Demographic data was collected through a survey[2] form shared with the customers. Each row represents a survey response from each individual.

### 2.1 Purchase Data

In the original Purchase data, there were 20 columns and 3,179 rows of data, with the first two columns being indexes. Data cleaning was then conducted to remove 13 duplicated rows and the two redundant index columns. For a detailed data description, refer to Appendix A.

There were also 175 mismatches between 66 unique *Invoice IDs* and variables *Order Date*, *Name*, *State* and *City* found in the dataset. This data anomaly implied that multiple customers have made purchases under one *Invoice ID*. Another alternative scenario is that a customer has purchased items months apart under the same *Invoice ID*. Both cases are illogical and these mismatched rows were removed from the dataset since more information with regards to the data collection process would be required to retain these records.

After further exploration, a time gap in the Purchase data was identified as there were no purchase records between 2012 to 2014. There were only 12 rows of data recorded in 2011 and these individuals have not made any purchases after 2011. These rows were then removed as they were deemed as outliers which would skew the subsequent customer segmentation analysis. With that, the purchase data now spans from April 2015 to December 2021.

After data cleaning and preparation, there are 2,969 rows of purchase data across 18 columns made by 293 unique customers.

### 2.2 Psychometric and Demographic Data

In the original Psychometric and Demographic data, there were 59 columns and 304 rows of data. Besides the *'Name'* column used to identify the respondent, the remaining columns represent responses to multiple-choice, rating and open-ended questions in the survey form. There are a total of 16 demographic questions and 42 psychometric questions, totalling up to 58 variables.

The psychometric questions can be sectioned into the following eight main aspects defined by the author of the dataset — *'Satisfaction with Life', 'Collectivism/Allocentrism', 'Individualism/Idiocentrism', 'Long-term Orientation', 'Short-term Orientation', 'Materialism', 'Spiritualism',* and *'Environmental Behaviour'*. Questions placed under the same section header can be interpreted to be collectively measuring the same psychometric need. For example, *'I am willing to give up today's fun for future success'* and *'I believe persistence is key to success'* variables belong under the *'Long-term Orientation'* psychometric aspect. For a detailed data description, refer to Appendix B.

[2] Google Survey Form (updated) created by the dataset authors can be accessed here: https://docs.google.com/forms/d/1Ncinn_BE_vdlrOuikhdkX3zsHPj1ZdRzG_3DIn0PtzI/edit

There were multiple data quality issues observed that required data cleaning and standardisation.

Firstly, there were some questions dependent on the previous response in the form. For example, the question '*If yes*' was asked twice as a follow-up to find out about the frequency of the related activity. More specifically, it appeared after the question '*Do you practise meditation?*' and '*Do you do any form of exercise*?'. Thus, these ambiguous columns were renamed to '*If you do meditate, how often do you practise meditation?*' and '*If you exercise, how often do you exercise?*' respectively. The related activity mentioned in the previous question was included by renaming the *'If Yes'* columns to a more intuitive name for analysis in the later stages.

Next, the response types were unstandardized due to the difference in nature of the questions and anomalies discovered within each question type. For rating questions that allowed respondents to only select from a scale of 1 to 7, the *'The things I own say a lot about how well I'm doing in life'* question had 10 instances with responses recorded as 4.5. These float values were then rounded to the nearest integer to standardise with the range of values in other questions.

For open-ended questions where respondents could input any response, there were multiple instances where the same word was represented in different formats due to differences in spelling, capitalisation or spaces. Relevant text preprocessing such as trimming whitespaces and capitalising the first letter were conducted. Furthermore, values with similar meanings were also grouped together. For example, in the '*Current Job Title*' column, '*Jobless*' and '*Unemployed*' were combined to represent one value instead.

Upon further exploration, there were also instances of duplicated and missing values. Duplicates could occur if a respondent submitted the survey form twice. The first occurrence was retained for a duplicated '*Name*' instance spanning across 2 entries. On the other hand, the '*Hobbies Count*' and '*Age*' columns contained a small number of missing or zero values and were filled with the mode and mean values, respectively.

After data cleaning and preparation, there are 303 rows of data across 59 columns.

## 3 Data Preparation

In this stage, the cleaned data is transformed into relevant variables for subsequent customer segmentation and analysis using domain knowledge.

Based on the Pareto Principle, 20% of customers account for 80% of the firm's profits (Prasad, 2019). In order to group customers into meaningful segments with common characteristics, both Recency, Frequency and Monetary Value (RFM) Segmentation and Needs-based Segmentation were considered.

Psychometric data representing customers' underlying needs could potentially be used to perform a Needs-based Segmentation. However, the dataset only contains Psychometric data for a particular point in time. It would be impossible to construct a Transition Matrix to understand how customers evolve over time, rendering the subsequent calculation of CLV to be impossible as well. Without understanding how customer segments evolve and their corresponding lifetime value, Flipkart would be unable to gain a better understanding of the financial returns of investing in different customers.

With these considerations in mind, **RFM Segmentation** was selected to be conducted with the Purchase data instead since it is possible to construct the Transition Matrix and calculate the CLV. Furthermore, it is a popular marketing technique that identifies the most valuable customers by analysing how recent the customer's last purchase was, how often they purchased and how much they spent in a particular time period.

### 3.1 Purchase Data

Calculating Recency, Frequency and Monetary Value

Using the dataset, *Recency* was calculated as the number of days between the customer's last purchase date and the overall latest date identified in the dataset (9 December 2021). For example, if the customer last purchased on 9 December 2021, the Recency value will be 0. Next, *Frequency* was calculated by summing up the total number of orders made by a customer. Finally, *Monetary Value* was derived by summing the *'Final Price'* variable, which represents the total amount for different products within each transaction, for all orders purchased by a customer.

Subsequently, customers were ranked within each RFM variable from 1 to 293. Lower R values indicate more recent purchases while higher F and M values indicate a higher frequency and value of purchases respectively. Accordingly, customers with lower R, higher F and M values are deemed more profitable to a company and thus are ranked with a higher value in each RFM variable respectively. Thereafter, the ranking of each customer within each RFM variable was normalised to a score out of 100.

### 3.2 Psychometric and Demographic Data

Creating Numerical Variables

For the subsequent stages of this project, the categorical variables were converted into numerical form to prepare them as input features for the classification model later on. One-hot encoding was used for binary variables such as *'Gender'*, *'Has Child'*, and *'Do you do any form of exercise?'*. Ordinal encoding was used for variables where there is a known relationship between the levels of categories. Some examples include *'Income (per month)'* and *'Exercise Frequency'*. Dummy encoding was used for the remaining nominal categorical variable *'Preferred Mode of Payment'* as it contained more than two categories. The dataset was expanded to 64 columns (including *'Name'* which will not be considered as a descriptor variable).

Factor Analysis

Due to the large number of columns present in the dataset, factor analysis was conducted on the Psychometric and Demographic data to reduce the data dimensionality. Data with a Kaiser-Meyer-Olkin score above 0.6 is usually deemed suitable for dimensionality reduction techniques like factor analysis, as higher scores indicate more correlation among variables that can potentially be reduced by grouping together (Babu, 2020).

While the data achieved a score slightly above 0.6, giving the green light for factor analysis to be conducted, the quality of factors was still less than ideal. As the Psychometric and Demographic variables did not exhibit high correlation among the variables, the factors distilled were unable to effectively capture much of the variation present in the data.

## 4 Customer Segmentation

As mentioned earlier, customer segmentation will be conducted using RFM segmentation.

### 4.1 Determining RFM Weights

To determine the overall RFM score, different techniques were explored to determine the weights assigned to each RFM variable.

In an e-commerce business, it is likely for a customer to search and purchase products regularly and frequently, resulting in a higher emphasis on Recency and Frequency variables over the Monetary Value variable (Makhija, 2021). Hence, in the calculation of the RFM score, higher weights should be given to the R and F scores compared to the M score.

Estimating RFM Weights with Linear Regression Models

Firstly, a preliminary linear regression model was executed to determine the importance of R, F and M in predicting whether a customer will purchase in the next period. The model returned the coefficients of (1.3361, 0.2431, 0) which will be normalised to (0.8461, 0.1539, 0) for R, F and M, respectively. However, the performance of the best model returned an R-squared score of 0.387, suggesting that the model only explains 38.7% of the total variance. With such lacklustre results, the coefficients of the RFM variables from the linear regression model will not be used to derive the weighted RFM score.

Setting RFM Weights Empirically

Instead, weights of (0.4, 0.4, 0.2) were empirically assigned to R, F and M values respectively, placing more importance on Recency and Frequency while not neglecting the importance of Monetary Value. The following formula was used to calculate the overall weighted RFM score for each customer:

$$\text{RFM score} = 0.4 * R + 0.4 * F + 0.2 * M$$

Based on the following histogram, the derived RFM scores for the customers were uniformly distributed between 0 to 100. The scores were then used to categorise the customers into numerous segments.

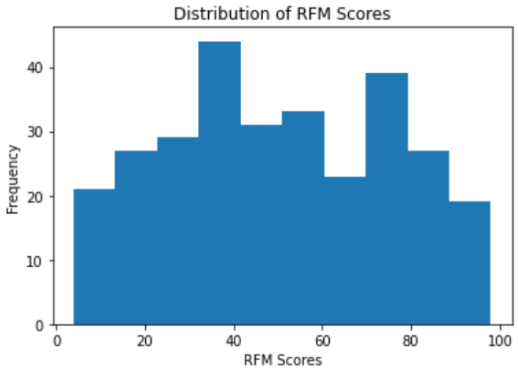



*Figure 1: Histogram of RFM Scores*

### 4.2 Determining an Ideal number of Segments

Although the optimal number of segments from RFM Segmentation tends to be determined by managerial judgement, K-Means Clustering on the RFM variables (Chugh, 2021) was also explored as an alternative approach by using the Elbow Method to find the number of segments. However, as the segments obtained from K-Means Clustering are unlabelled, it may not be as accurate to analyse the segments without inputs from managers and experts who have a good understanding of the business in Flipkart. Hence, to avoid introducing greater uncertainty and subjectivity, **RFM Segmentation using different cut-off RFM scores** was adopted as it is more commonly used in the industry.

To identify the best number of segments for a relatively small dataset, managerial judgement together with numerical and strategic criteria were involved in the decision process. The customers were grouped into 2, 3, 4 and 5 segments where each respective segmented group was labelled intuitively based on the RFM scores. The breakdown and profile of the various segments are as follows:

| Customer Segment | RFM Score | Count | Percent | Recency | Frequency | Monetary Value |
|---|---|---|---|---|---|---|
| High Value | 51-100 | 141 | 48.1% | 112 | 17 | 30,782 |
| Low Value | 0-50 | 152 | 51.9% | 344 | 4 | 10,134 |

*Table 1: 2 Segments*

The two segments are clearly defined with large differences observed in the RFM values. However, these segments may be too generic for Flipkart to craft effective targeted marketing campaigns in subsequent stages. Thus, it will be ideal to further split the customers into more segments to minimise the differences present within each segment.

| Customer Segment | RFM Score | Count | Percent | Recency | Frequency | Monetary Value |
|---|---|---|---|---|---|---|
| High Value | 76-100 | 61 | 20.8% | 66 | 26 | 40,311 |
| Medium Value | 51-75 | 80 | 27.3% | 147 | 10 | 23,516 |
| Low Value | 0-50 | 152 | 51.9% | 344 | 4 | 10,134 |

*Table 2: 3 Segments*

| Customer Segment | RFM Score | Count | Percent | Recency | Frequency | Monetary Value |
|---|---|---|---|---|---|---|
| High Value | 76-100 | 61 | 20.8% | 66 | 26 | 40,311 |
| Medium Value | 51-75 | 80 | 27.3% | 147 | 10 | 23,516 |
| Low Value | 26-50 | 98 | 33.4% | 255 | 5 | 11,555 |
| Lost | 0-25 | 54 | 18.4% | 506 | 2 | 7,555 |

*Table 3: 4 Segments*

As shown from the average RFM values, both 3 and 4 Segments have the same characteristics for High Value and Medium Value customers due to the same ranges of RFM scores defined. The

only difference is that the Low Value customers in 3 Segments were further broken down into Low Value and Lost customers in 4 Segments. Since Lost customers are generally deemed unprofitable to companies, 4 Segments is preferred over 3 Segments as Flipkart would be able to redirect its resources that were intended for Lost customers to the other more profitable customer segments.

| Customer Segment | RFM Score | Count | Percent | Recency | Frequency | Monetary Value |
|---|---|---|---|---|---|---|
| Top | 81-100 | 41 | 14.0% | 57 | 30 | 46,988 |
| High Value | 61-80 | 67 | 22.9% | 110 | 14 | 28,881 |
| Medium Value | 41-60 | 69 | 23.5% | 198 | 7 | 13,734 |
| Low Value | 21-40 | 73 | 24.9% | 298 | 4 | 10,181 |
| Lost | 0-20 | 43 | 14.7% | 533 | 2 | 7,635 |

*Table 4: 5 Segments*

Across five segments, the customers are evenly distributed and each segment is mostly well-represented. However, given the relatively small dataset, the segments could potentially be overfitted and too refined, which may not be representative of the population. Furthermore, having more segments would likely incur greater marketing costs as more resources are required to design a greater number of targeted marketing strategies and loss of economies of scale.

With these considerations in mind, **4 Segments** was chosen as the optimal number of segments.

## 5 Transition Matrix

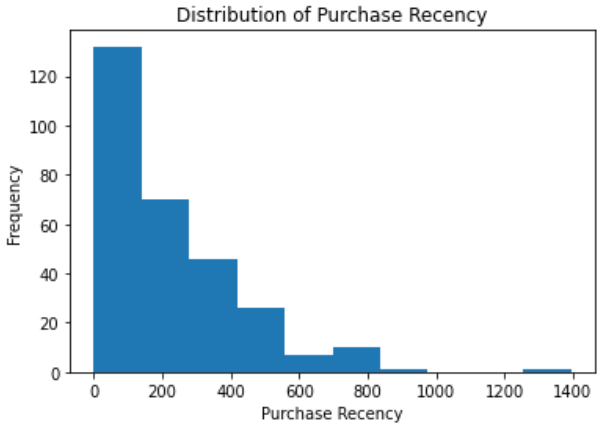



*Figure 2: Histogram of Purchase Recency*

A Transition Matrix is required to measure how customers have evolved from one segment to another for a particular time period. To define an appropriate time period, the distribution of the *Recency* values was analysed from the histogram plotted above, where most values were within a year (365 days). To ensure a sizable number of customers were kept for analysing how customers evolved, a time period of **three months or quarterly** was deemed the most appropriate. In

addition, as an e-commerce company that sells mostly fast-moving consumer goods, analysing how customers evolve in a shorter time period would be more suitable.

Next, there is a need to identify the segments of the customers in the **last quarter** to analyse how the customers move across segments between the current and previous period. As the latest transaction within the Purchase Data was 9 December 2021, the data was split into transactions before and after 9 September 2021 (3 months earlier).

A total of 5 new customers who purchased after 9 September 2021 were removed as there were no prior records from these customers before that date. Thus, they are excluded from the derivation of the transition matrix. The remaining 288 customers were segmented into the 4 Segments as concluded earlier. The following transition matrix was then obtained by tabulating the proportion of customers that move across segments before and after 9 September 2021.

| **Customer Segment** Last Period/Next Period | High Value | Medium Value | Low Value | Lost |
|---|---|---|---|---|
| High Value | 0.6909 | 0.3091 | 0 | 0 |
| Medium Value | 0.2353 | 0.6235 | 0.1412 | 0 |
| Low Value | 0.0323 | 0.0968 | 0.8172 | 0.0538 |
| Lost | 0 | 0.0182 | 0.0909 | 0.8909 |

*Table 5: Transition Matrix of Customer Segments*

## 6 Customer Lifetime Value

To calculate the revenue of customers within the specified time period, the dataset was filtered to include only transactions that occurred within the last quarter (9 September to 9 December 2021). The average revenue for each segment in the last quarter was then obtained.

Next, future customers in each segment were predicted for the following five periods using the existing number of customers and the transition matrix created previously. Assuming that the transition matrix remains constant over time, the evolution of customers across segments in the future can be predicted accurately. The quarterly revenue for the segments was calculated and summed cumulatively over the six time periods which included the current and predicted future five periods.

Thereafter, an annual discount rate of 25% was used to discount the cumulated revenues to reflect the net present value as of the current period. Converting the annual discount rate into quarterly compounding rates is equivalent to 5.74% quarterly. This relatively high discount rate was chosen to remain focused on the short-term revenues due to the expected slowdown in the Indian economy in 2022 (The Economic Times, 2022).

The resulting value at the end of the sixth time period reflects the total CLV for each customer segment. However, the total CLV could lead to misleading interpretations if some segments contain a larger number of customers relative to other segments and result in a higher CLV simply due to more customers. As such, the total CLV was divided by the initial number of customers within each segment to obtain the average CLV per customer instead, which is a more

accurate measure of the value for each segment. From the figures below, it can be observed that the High Value customers generate the highest average CLV while the Lost customers do not generate any CLV.

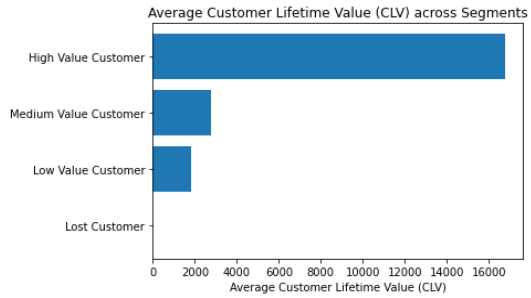



*Figure 3: Bar Chart of Average CLV across Segments*

| Customer Segment | Count | Average CLV | Total CLV |
|---|---|---|---|
| High Value | 61 | 16,781 | 1,023,670 |
| Medium Value | 80 | 2,794 | 223,535 |
| Low Value | 98 | 1,836 | 179,886 |
| Lost | 54 | 0 | 0 |
| **Total** | **293** | - | **1,427,091** |

*Figure 4: Average and Total CLV Breakdown*

The total Customer Base Lifetime Value for the 293 customers in the sample dataset amounts to 1.427 million INR. Flipkart has a registered customer base of more than 350 million (Flipkart, 2021). Assuming each sample customer in the dataset represents approximately 1,194,539 (350 million divided by 293) Flipkart customers in the population, it can be approximated that Flipkart's Customer Base Lifetime Value for the current and the next five periods is worth 1.705 trillion INR.

## 7 Segment and Descriptor Analysis

Segment analysis was conducted on the segmentation and descriptor variables separately. One-Sample t-Test was used to determine whether the centroid of each respective segment is statistically different from the population mean, represented by colour blocks in green (higher) or red (lower). It is assumed that the sample mean from 293 customers within the dataset approximates the population mean of the entire Flipkart customer base.

In order to effectively identify the unique characteristics of each segment, it is more insightful to only evaluate variables that are statistically different from the rest of the population. Thus, only statistically significant variables are retained for this analysis.

### 7.1 Segment Analysis

| Segmentation Variable | Population | Customer Segment | | | |
|---|---|---|---|---|---|
| | | High Value | Medium Value | Low Value | Lost |
| Recency | 232.44 | 64.98 | 147.13 | 253.19 | 505.38 |
| Frequency | 10.13 | 26.26 | 10.43 | 4.55 | 2.12 |
| Monetary Value | 20,070.56 | 41,731.24 | 23,785.42 | 11,387.67 | 7,805.17 |

Lower ($p<0.05$)
Lower ($p<0.10$)
n.s.
Higher ($p<0.10$)
Higher ($p<0.05$)

*Figure 5: Segmentation Variables Differences per Segment*

It is observed that for each segment, there is at least one segmentation variable statistically different from the rest of the population, forming a clear divide between segments. **High Value customers** tend to make high-value frequent purchases and most likely have made a recent

purchase. **Medium Value customers** would have made a recent purchase, but not as recent as High Value customers. **Low Value customers** tend to make occasional low-value purchases. **Lost customers** have last made a purchase quite some time ago and make lower-value purchases less frequent than Low Value customers.

## 7.2 Descriptor Analysis

| Descriptor Variable | Population | Customer Segment | | | |
|---|---|---|---|---|---|
| | | High Value | Medium Value | Low Value | Lost |
| Has Child | 0.14 | 0.12 | 0.06 | 0.21 | 0.08 |
| Meditation | 0.45 | 0.36 | 0.41 | 0.55 | 0.48 |
| Age | 26.15 | 26.74 | 25.37 | 27.31 | 23.88 |
| Care for Chronic Illness | 0.14 | 0.09 | 0.06 | 0.18 | 0.21 |
| Preferred Mode of Payment - Credit Card | 0.07 | 0.17 | 0.04 | 0.07 | 0 |
| Preferred Mode of Payment - Net Banking | 0.03 | 0.02 | 0 | 0.03 | 0.06 |
| Preferred Mode of Payment - Online Wallets | 0.02 | 0 | 0.01 | 0.02 | 0.08 |
| If I could live my life over, I would change almost nothing | 4.21 | 4.64 | 4.24 | 4.11 | 3.9 |
| I feel good when I co-operate with others | 6.02 | 5.97 | 6.2 | 5.85 | 6.15 |
| When making a decision, I take other people's needs and feelings into account. | 5.62 | 5.74 | 5.82 | 5.4 | 5.54 |
| It is my duty to take care of my family, even when I have to sacrifice what I want. | 5.96 | 5.83 | 6.19 | 5.9 | 5.75 |
| It upsets me when my work is not recognized by others | 5.11 | 5.14 | 5.03 | 5.01 | 5.46 |
| I believe success in life does not mean becoming rich | 5.56 | 5.43 | 5.61 | 5.8 | 5.19 |
| I walk/cycle/use public transport to save fuel | 5 | 4.66 | 4.71 | 5.33 | 5.04 |

Lower ($p<0.05$)
Lower ($p<0.10$)
n.s.
Higher ($p<0.10$)
Higher ($p<0.05$)

*Figure 6: Descriptor Variables Differences per Segment*

Out of all 63 descriptor variables, 14 of them are statistically different from the rest of the population. Analysis will be carried out with reference to the 8 main psychometric aspects detailed in Appendix B.2.

**High Value customers** tend to prefer Credit Cards as a mode of payment and dislike using Online Wallets. These individuals have a higher life satisfaction as they indicate on a higher scale that they would change almost nothing if they could live their lives over. With increased spending on credit cards, benefits such as cashback and reward points can be earned. Studies have suggested that credit cards motivate spending by exploiting reward networks in the brain which encourages spending habits and even shopping addictions (Prelec, 2021). This payment method preference for Credit Cards could be closely associated with the higher Frequency and Monetary Value of High Value customers.

**Medium Value customers** tend to have no children or the need to care for individuals with chronic illnesses in comparison to other segments. They do not prefer Net Banking as a mode of payment. These customers also have higher collectivism and allocentrism as they have rated related questions under this aspect significantly higher than the population. More specifically, they stated that they take into account others' needs and feelings during decision-making. Furthermore, they enjoy cooperating with others and even make personal sacrifices to take care of their family. An interesting pattern observed is that Medium Value customers indicated their

willingness to take care of family members, but they tend to not have children or family members with chronic illnesses which require additional care and attention.

**Low Value customers** as compared to other segments tend to have children and more of them practise meditation. They are rather eco-friendly as they take extra efforts to save fuel through means such as walking, cycling or taking public transport. Furthermore, the Low Value customers also have a more spiritual outlook in life where they believe success does not solely depend on wealth.

**Lost customers** are significantly younger than the rest of the population and do not prefer using Credit Cards as a mode of payment. They are less spiritual as they strongly believe that success in life means becoming rich, as opposed to Low Value customers. Furthermore, they have stronger individualism as being recognised by others is important to them and they get upset if that does not happen.

## 8 Classification Model

Predicting the segment membership for new customers will be insightful for Flipkart to better tailor marketing campaigns to acquire and retain these individuals. Descriptor variables may be more easily obtained as compared to segmentation variables. Thus, this section explores whether descriptors could be used to predict segment membership using a multinomial logistic regression model.

Since psychometric information is usually obtainable only when new customers fill in surveys, it may not always be attainable. Thus, three different models were explored with demographic, psychometric or both, as the descriptor variables respectively. Grid search was used for hyperparameter tuning and 5-fold cross validation was used to evaluate the models. Various data transformation techniques such as Box-Cox transformation, Log transformation and Min-Max Scaler were also applied to these models. The best model results from the variations tested are as follows:

| Model | Test | | Validation | |
|---|---|---|---|---|
| | Accuracy | Balanced Accuracy | Accuracy | Balanced Accuracy |
| Demographic | 0.2807 | 0.2500 | **0.3379** | **0.2457** |
| Psychometric | 0.3158 | 0.2921 | 0.3332 | 0.2401 |
| Psychometric and Demographic | **0.3333** | **0.3077** | 0.3332 | 0.2363 |

*Table 6: Descriptor Variable Classification Model*

In a 4-segment setting, a model is deemed to be better than randomly selecting a segment if the accuracy is higher than 0.25. However, it has been observed that the model has predicted all test entries as the majority class on a few occasions. Unlike accuracy which predicts well on the majority class, balanced accuracy ensures that the prediction of both minority and majority classes in the model are accounted for (Grandini et al., 2020). Hence, to gain a better perspective

across the four imbalanced segments defined, balanced accuracy is a superior metric for evaluation.

The best performing model would be the psychometric demographic model but there is still significant room for improvement with the balanced test accuracy at 0.3077. Without psychometric data available, the demographic model can only achieve performance as good as a random guess. Hence, instead of relying only on the more easily attainable demographic data, Flipkart could survey its customers to gain a stronger understanding of their needs which would help in predicting the segments of future customers.

## 9 Marketing Strategies and Recommendations

For Flipkart to determine whether to conduct the following marketing campaigns, the marketing costs involved to acquire or retain customers should not exceed the expected average CLV within each segment. While it is important to attract new customers and gain more market share, there could be high acquisition costs associated with increased spending on promotional advertisements and higher likelihood of return and exchange rates from new customers. Thus, for Flipkart to remain competitive in the e-commerce market, it would be more cost-effective and worthwhile to drive repeat sales to current customers by building long-term customer relationships and maximising its customer retention rates. As such, the following marketing strategies and recommendations are tailored to each customer segment, excluding Lost customers, in order to effectively target them.

### 9.1 High Value Customers

Using Textual Analytics to Offer Wide Product Selection

With their high satisfaction in life as indicated by related questions, the underlying needs of High Value customers would be to maintain the same level of satisfaction and quality of life. Based on research conducted by Leelanuithanit et al. (1991), factors such as material possession have a significant positive impact on overall life satisfaction. Thus, Flipkart can aim to meet these customers' needs of maintaining their life satisfaction in this aspect by providing tailored product offerings that best suit their current lifestyles.

Currently, Flipkart has a high volume and range of products that include both higher-end popular brands such as Apple and in-house brands such as Citron. The wide selection of brands and products provides numerous options that can cater to the individual preferences of each customer. Flipkart was projected to have 420,000 sellers by the end of 2021 (Flipkart, 2021). However, Flipkart's main competitor, Amazon India, has more than 1 million sellers on its platform (The Economic Times, 2021).

To increase the variety of sellers and products, Flipkart could make use of textual analytics to identify products that customers search for on its platform. If there is a lack of sellers for a particular product, Flipkart could try to acquire more of such sellers, which will encourage more High Value customers to purchase on its platform to meet their needs.

Enhancing Credit Card Promotions

High Value customers were observed to have a stronger preference for credit cards as a payment method. Currently, many leading Indian banks have existing partnerships with Flipkart to offer attractive discounts and cashback offers on purchases with credit cards (Bank Bazaar, n.d.). In particular, the Flipkart Axis Bank Credit Card allows consumers to earn additional cashback on Flipkart, higher than any other credit card (Axis Bank, n.d.).

With consumers signing up for the Flipkart credit card, they will be more likely to make their purchase from Flipkart as opposed to its competitors without any benefits. Flipkart could disseminate credit card promotional materials to High Value customers that prefer this payment mode, further encouraging purchases from them.

### 9.2 Medium Value Customers

Fine Tuning Advertisements Efforts


*Figure 7: Current Flipkart Advertisement (Flipkart, n.d.)*


*Figure 8: Suggested Flipkart Advertisement (Marketing Interactive, 2021)*

Medium Value customers were identified to have higher collectivism and allocentrism. Flipkart can make use of this insight to reduce advertising costs through developing personalised advertising campaigns on its website. For instance, Flipkart can use pronouns such as "we" instead of "I" or display promotional advertisements of people in groups as seen in Figure 8 (Nahai, 2013) instead of its current approach of displaying only individuals (Figure 7). Both strategies applied to an advertisement's text and graphics will introduce the idea of inclusiveness, which could engage Medium Value customers better. By tailoring the user's experience on its platform, this builds brand engagement with Medium Value customers, thus increasing likelihood of purchase from them.

### 9.3 Low Value Customers

Analysing Purchase Patterns and Optimising User Experience

Low Value customers are on average less than 30 years old and are more likely to have children as compared to the population. As young millennial parents who are digitally savvy and have busy schedules on top of taking care of their children, e-commerce shopping provides great convenience and flexibility that enables them to buy necessities with ease. However, parents have a clear idea of the products they are seeking and seldom look for product variety (MikMak, n.d.). Coupled with the extra urgency to cater to their children's needs, these customers would easily switch to other competitor sites should the product be out of stock or the transaction process is cumbersome.

Flipkart can analyse the purchase patterns to understand the popular products these customers are looking for and push more accurate product recommendations that can capture their interest. Additionally, the website features can be further optimised to ensure a good user experience that can provide a seamless purchase transaction. Enhanced page navigation featuring relevant products or sales and higher load speeds can help customers find their ideal products more easily, saving additional time or hassle.

Having a Robust Customer Support

Low Value customers are more spiritual based on their responses to spiritualism-related questions. Flipkart can adopt spiritual marketing by being clear about what they have to offer and focusing on resolving problems for Low Value consumers. According to Jain (2020), spiritual consumers tend to instantly know whether their needs can be met by companies. As such, Flipkart can consider dedicating more resources to its Customer Support Twitter account[3] to resolve every customer's enquiry as soon as possible. According to Porter (n.d.), people who tweet negatively still feel more favourable towards companies who respond to their concerns which increases the likelihood of purchase at these companies. Thus, by being customer-centric through resolving any purchase-related issues, Flipkart can build customer loyalty and increase revenue with minimal cost.

### 9.4 Lost Customers

Lost customers are not expected to generate any future revenue for the company and thus, Flipkart should not target these customers.

## 10 Limitations

There were a few limitations to this study associated with the usage of secondary data.

The dataset author has provided limited data documentation and **uncertainty regarding the data collection process** remains. To derive a sample dataset representative of the population, respondents should be randomly sampled. However, it is unclear how the customers within the dataset were selected.

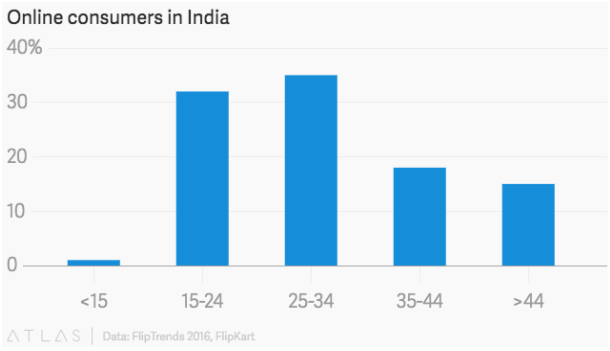



*Figure 9: Bar Plot of Flipkart Customer Age (Bhattacharya, 2016)*

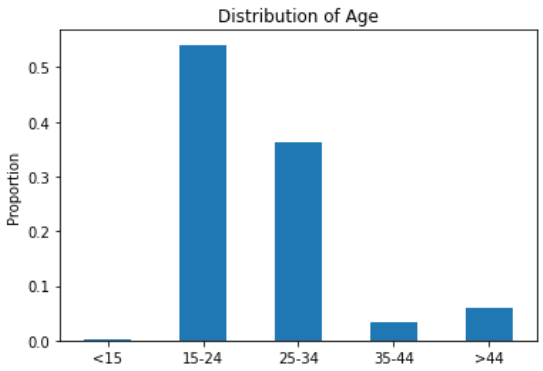



*Figure 10: Bar Plot of Sample Data Customer Age*

[3] Flipkart's Official Customer Support Twitter Account https://twitter.com/flipkartsupport

Comparing the distribution of customers' age within the dataset against Flipkart population data as seen in the figures above, differences can be observed. A large proportion of customers from the dataset are between the age of 15 to 34. The dataset author may have collected the Flipkart data within his community, consisting of similar and like-minded individuals where a higher proportion of respondents belonged to the 15 to 24 age group, causing sampling bias. More specifically, non-response or voluntary response bias could have been embedded in the collection process. Customers who are willing to provide their transaction history and answer the survey may have underlying latent differences compared to customers not observed in the sample dataset.

Since information with regards to these concerns is unavailable, it is unsure whether the dataset collected is entirely free from bias. Despite that, the methodologies adopted in this project are robust and free from biases as much as possible. Thus, conclusions made from this dataset could remain insightful for Flipkart.

Another implication of this dataset is the **limited size** with only 303 unique customers. A small dataset makes it difficult to extrapolate the findings discussed to the general population of Flipkart customers, especially in cases where outliers are present. Augmenting the adopted methodology in this project with more data could further generate more meaningful insights for Flipkart.

## 11 Summary

Through this project, RFM customer segmentation was performed to group customers into meaningful segments which were then evaluated. This allows Flipkart to identify distinct customer segments and target them through effective strategies crafted based on their characteristics to subsequently capture and maintain valuable long-term customer relationships. However, further room for improvement exists to use higher quality data to generate more meaningful insights which can better reflect Flipkart's customer base.

## References


Axis Bank. (n.d.). *Flipkart Axis Bank Credit Card - 5% Cashback on Flipkart, Myntra*. Axis Bank. Retrieved April 14, 2022, from https://www.axisbank.com/retail/cards/credit-card/flipkart-axisbank-credit-card/features-benefits

Babu, D. (2020, October 25). *Dimensionality Reduction using Factor Analysis (Python Implementation)*. Analytics Vidhya. Retrieved April 14, 2022, from https://www.analyticsvidhya.com/blog/2020/10/dimensionality-reduction-using-factor-analysis-in-python/

Bank Bazaar. (n.d.). *Flipkart Credit Card Offers Check - Promo's, Deals as on 13 Apr 2022*. BankBazaar. Retrieved April 14, 2022, from https://www.bankbazaar.com/flipkart-credit-card-offers.html

Bhattacharya, A. (2016, December 27). *A typical online shopper in India is a man aged 25-34 buying electronics through his mobile phone*. Quartz. Retrieved April 14, 2022, from https://qz.com/india/872834/an-average-online-shopper-in-india-is-a-man-aged-25-34-years-buying-electronics-through-his-mobile-phone/

Chugh, J. (2021, April 3). *Using K-means to segment customers based on RFM Variables*. Medium. Retrieved April 14, 2022, from https://medium.com/web-mining-is688-spring-2021/using-k-means-to-segment-customers-based-on-rfm-variables-9d4d683688c8

Dayalani, V. (2021, December 21). *Amazon Vs Flipkart: Who Led The Indian Ecommerce War In 2021?* Inc42. Retrieved April 12, 2022, from https://inc42.com/datalab/amazon-vs-flipkart-who-led-the-indian-ecommerce-war-in-2021/

The Economic Times. (2021, December 15). *amazon sales: Amazon crosses 10 lakh sellers mark in India*. The Economic Times. Retrieved April 14, 2022, from https://economictimes.indiatimes.com/industry/services/retail/amazon-crosses-10-lakh-sellers-mark-in-india/articleshow/88304944.cms?from=mdr

Flipkart. (n.d.). Online Shopping Site for Mobiles, Electronics, Furniture, Grocery, Lifestyle, Books & More. Best Offers! Retrieved April 14, 2022, from https://www.flipkart.com/

Flipkart. (2021, September 16). *Flipkart on-track to have over 4.2 lakh sellers & MSMEs by December 2021*. Flipkart Media. Retrieved April 14, 2022, from https://storiesflistgv2.blob.core.windows.net/stories/2021/09/16092021-Press-Release-Flipkart-on-track-to-have-over-4.2-lakh-sellers-MSMEs-by-December-2021.pdf

Flipkart. (2021, October 28). *Flipkart and Moj announce a collaboration for Video and Live Commerce*. Final draft 27.10 | Press Release_Flipkart X Moj Partnership. Retrieved April 14, 2022, from https://storiesflistgv2.blob.core.windows.net/stories/2021/10/Press-release-Flipkart-and-Moj-announce-a-collaboration-for-Video-and-Live-Commerce.pdf

Grandini, M., Bagli, E., & Visani, G. (2020, August 14). Metrics for Multi-Class Classification: an Overview. https://doi.org/10.48550/arXiv.2008.05756

Jain, P. (2020, August 30). Spirituality and Marketing: An Amalgam. Retrieved April 14, 2022, from https://www.linkedin.com/pulse/spirituality-marketing-amalgam-praggya-jain/

Khatri, B. (2019, December 18). *Why Is Flipkart Making Losses Despite Revenue Growth?* Inc42. Retrieved April 12, 2022, from https://inc42.com/features/why-is-flipkart-making-losses-despite-revenue-growth/

Leelanuithanit, O., Day, R., & Walters, R. (1991, June 1). Investigating the Relationship between Marketing and Overall Satisfaction with Life in a Developing Country. *Journal of Macromarketing*, *11*(1). https://journals.sagepub.com/doi/10.1177/027614679101100102

Makhija, P. (2021, June 3). *RFM Analysis for Customer Segmentation*. CleverTap. Retrieved April 13, 2022, from https://clevertap.com/blog/rfm-analysis/

Marketing Interactive. (2021, October 15). *DFI Retail Group launches mega supermarket at Sai Wan*. Marketing Interactive. Retrieved April 14, 2022, from https://www.marketing-interactive.com/dfi-retail-group-launches-mega-supermarket-at-sai-wan

MikMak. (n.d.). *4 Ways Parents are Buying Online for (and with!) Their Kids*. MikMak. Retrieved April 14, 2022, from https://www.mikmak.com/blog/4-ways-parents-are-buying-online-for-and-with-kids

Nahai, N. (2013, July 15). *How to Sell Online to Individualist vs Collectivist Cultures*. Psychology Today. Retrieved April 14, 2022, from https://www.psychologytoday.com/us/blog/webs-influence/201307/how-sell-online-individualist-vs-collectivist-cultures

Poojary, T., & Ranjan, S. (2019, November 6). *Ecommerce paradox: Flipkart and Amazon India continue to make losses, but there's a silver lining*. YourStory. Retrieved April 13, 2022, from https://yourstory.com/2019/11/flipkart-amazon-india-losses-ecommerce-walmart/amp

Porter, S. (n.d.). *How Twitter has become a key customer support channel*. Twitter for Business. Retrieved April 14, 2022, from https://business.twitter.com/en/blog/how-twitter-has-become-a-key-customer-support-channel.html

Prasad, A. (2019, October 22). *Customer Segmentation with RFM analysis — Part 1 | by Aman Prasad | Merino Services Analytics Blog*. Medium. Retrieved April 13, 2022, from https://medium.com/merino-services-analytics-blog/customer-segmentation-with-rfm-analysis-part-1-8deadb046113

Prelec, D. (2021, June 9). *How credit cards activate the reward center of our brains and drive spending*. MIT Sloan. Retrieved April 13, 2022, from

https://mitsloan.mit.edu/experts/how-credit-cards-activate-reward-center-our-brains-and-drive-spending

Sahni, A. (2019, October 18). *What Is Flipkart's Business Model?* Inc42. Retrieved April 12, 2022, from https://inc42.com/features/what-is-flipkarts-business-model/

Shrivastava, A., Chanchani, M., & Sajjad, M. (2017, January 2). *To stay ahead in online retail race, Flipkart lost Rs 14 crore per day in FY16*. The Economic Times. Retrieved April 13, 2022, from https://economictimes.indiatimes.com/small-biz/startups/to-stay-ahead-in-online-retail-race-flipkart-lost-rs-14-crore-per-day-in-fy16/articleshow/56282351.cms

## Appendix

### Appendix A: Purchase Data Description

The purchase data that was obtained is detailed as follows. It should be noted that the dataset author has collected the dataset for a different purpose, which could explain data columns not used within this project.

| | Invoice ID | Name | Order Date | State | City | Categories | Subcategories | Ratings | Quantity | MRP | Final Price | GST% | City_Tier | Discount% | Delivery Fee% | Brand | Sale | Covid |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | OD103719706054443200 | G3M1R1 | 2015-08-23 | Delhi | New Delhi | Home & Kitchen | Kitchen Appliances | 4.1 | 1 | 1370.33 | 1545.0 | 18% | Tier_1 | 0 | 13 | Nova | No | No |
| 1 | OD106420064045076000 | G3M1R1 | 2016-07-01 | Delhi | New Delhi | Health & Personal Care Appliances | Health Care | 4.3 | 1 | 1164.76 | 1260.0 | 5% | Tier_1 | 0 | 9 | Omron | No | No |
| 2 | OD106420064045076001 | G3M1R1 | 2016-07-01 | Delhi | New Delhi | Clothing and Accessories | Books | 4.5 | 1 | 596.67 | 730.0 | 5% | Tier_1 | 0 | 23 | Johnson | No | No |
| 3 | OD109711988579254000 | G3M1R2 | 2017-07-17 | Delhi | New Delhi | Exercise & Fitness | Fitness Accessories | 4.3 | 1 | 199.00 | 199.0 | 18% | Tier_1 | 0 | 0 | HAANS | No | No |
| 4 | OD109711988579254000 | G3M1R2 | 2017-07-18 | Delhi | New Delhi | Health Care | Health Supplements | 4.1 | 1 | 4162.00 | 4162.0 | 28% | Tier_1 | 0 | 0 | Muscletech | No | No |

*Figure 11: Purchase Data Sample Rows*

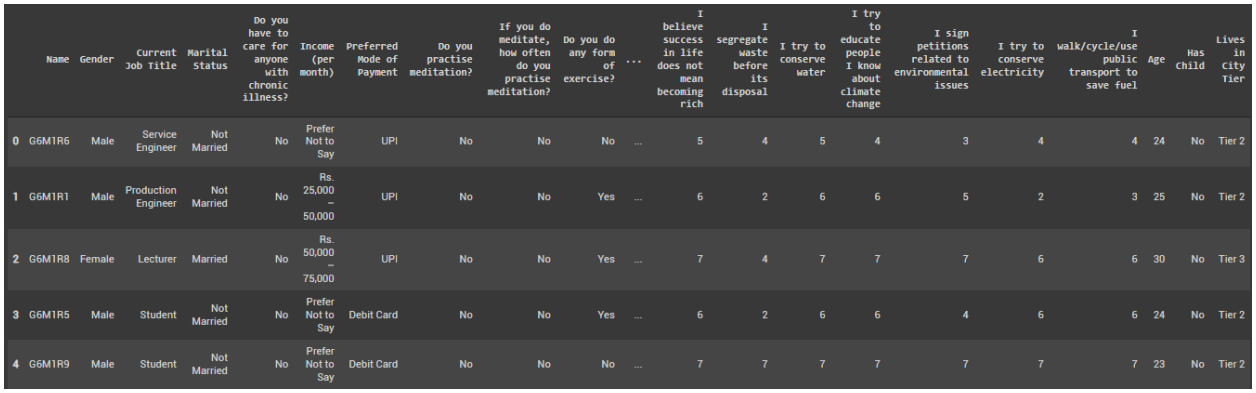
| **Column** | **Description** |
|---|---|
| Invoice ID | Unique code used to identify an invoice, which may contain multiple items purchased. |
| Name | Unique code used to identify a customer. |
| Order Date | Date order was placed. |
| State | State that customer lives in. |
| City | City that customer lives in. |
| Categories | Category of item purchased. |
| Subcategories | Subcategory of item purchased. |
| Ratings | Rating of items purchased between a scale of 1 to 5, made by other purchasers. |
| Quantity | Quantity of item purchased. |
| MRP | Maximum retail price for item purchased. |
| Final Price | Unit price of item purchased after discount, delivery fee, and tax. |
| GST% | Tax percentage for item purchased, categorised into 5 bins. |
| Discount% | Discount percentage for item purchased. |
| Delivery Fee% | Delivery fee percentage for item purchased. |
| Brand | Brand of item purchased. |

| | |
|---|---|
| City Tier | Tier of City, categorised into 3 tiers defined by the Indian government, based on the population density. |
| Sale | Indication of Flipkart sale period. |
| Covid | COVID wave based on time period and order date. |

*Table 7: Purchase Data Description*

**Appendix B: Psychometric and Demographic Data Description**

The psychometric and demographic data obtained is detailed as follows. Similarly, as the dataset author has collected the dataset for a different purpose, there may be questions which do not entirely fit within the context of our project.



| | Name | Gender | Current Job Title | Marital Status | Do you have to care for anyone with chronic illness? | Income (per month) | Preferred Mode of Payment | Do you practise meditation? | If you do meditate, how often do you practise meditation? | Do you do any form of exercise? | ... | I believe success in life does not mean becoming rich | I segregate waste before its disposal | I try to conserve water | I try to educate people I know about climate change | I sign petitions related to environmental issues | I try to conserve electricity | I walk/cycle/use public transport to save fuel | Age | Has child | Lives in city Tier |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | G6M1R6 | Male | Service Engineer | Not Married | No | Prefer Not to Say | UPI | No | No | No | ... | 5 | 4 | 5 | 4 | 3 | 4 | 4 | 24 | No | Tier 2 |
| 1 | G6M1R1 | Male | Production Engineer | Not Married | No | Rs. 25,000 – 50,000 | UPI | No | No | Yes | ... | 6 | 2 | 6 | 6 | 5 | 2 | 3 | 25 | No | Tier 2 |
| 2 | G6M1R8 | Female | Lecturer | Married | No | Rs. 50,000 – 75,000 | UPI | No | No | Yes | ... | 7 | 4 | 7 | 7 | 7 | 6 | 6 | 30 | No | Tier 3 |
| 3 | G6M1R5 | Male | Student | Not Married | No | Prefer Not to Say | Debit Card | No | No | Yes | ... | 6 | 2 | 6 | 6 | 4 | 6 | 6 | 24 | No | Tier 2 |
| 4 | G6M1R9 | Male | Student | Not Married | No | Prefer Not to Say | Debit Card | No | No | No | ... | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 23 | No | Tier 2 |

*Figure 12: Psychometric and Demographic Data Sample Rows*

Appendix B.1: Demographic Data Description

| **Demographic Questions** | **Description** |
|---|---|
| Gender | Indicates whether the respondent is Male or Female. |
| Age | Age of respondent. |
| Marital Status | Indicates whether the respondent is Single or Married. |
| Has Child | Indicates whether the respondent has a child or not. |
| Current Job Title | Occupation of the respondent, open-ended response. |
| Income (per month) | Income of the respondent, selected out of 7 pre-defined ranges. |
| Preferred Mode of Payment | Preferred payment method, selected out of 6 pre-defined options. |
| Lives in City Tier | Based on City Tier defined in respondent's latest purchase made. |
| Hobbies Count | Number of hobbies the respondent has. |
| How often do you pursue your hobbies? | Frequency of hobbies, selected out of 5 pre-defined options. Possible to input an open-ended response. |
| Do you practise meditation? | Indicates whether the respondent meditates or not. |
| If you do meditate, how often do | Frequency of medication, selected out of 5 pre-defined |

| | |
|---|---|
| you practise meditation? | options. Possible to input an open-ended response. |
| Do you do any form of exercise? | Indicates whether the respondent exercises or not. |
| If you exercise, how often do you exercise? | Frequency of exercise, selected out of 5 pre-defined options. Possible to input an open-ended response. |
| How often do you volunteer for social-service activities? | Frequency of volunteering, selected out of 5 pre-defined options. Possible to input an open-ended response. |
| Do you have to care for anyone with chronic illness? | Indicates whether the respondent has to take care of immuno-compromised individuals or not. |

*Table 8: Demographic Data Description*

Appendix B.2: Psychometric Data Description

The psychometric questions were based on the same scale where respondents could respond with an integer ranging from 1 to 7. Higher values are associated with greater belief in a specific scenario. These psychometric questions were categorised under the following eight psychometric aspects which are described as follows.

| **Psychometric Aspect** | **Description** |
|---|---|
| Satisfaction with Life | Emotions and feelings about one's directions and options for the future. |
| Collectivism/Allocentrism | Collectivistic mindset and beliefs with emphasis on cohesiveness and prioritisation of the group over the individual. |
| Individualism/Idiocentrism | Principle of being independent, autonomous, and self-reliant. Generally motivated by own preferences, rather than group goals and social norms. |
| Long-term Orientation | Focus on the future, placing emphasis and value on persistence, perseverance, and adaptability. |
| Short-term Orientation | Focus on the present, stronger value placed on immediate gratification than long-term fulfilment. |
| Materialism | Concern for possessions or material wealth and physical comfort, as opposed to spiritual or intellectual pursuits. |
| Spiritualism | Belief in immaterial and intangible reality that cannot be perceived by the senses, beyond the reach of purely materialistic interpretations. |
| Environmental Behaviour | Engagement in more environmentally conscious behaviour aimed at avoiding harm to and/or safeguarding the environment. |

*Table 9: Psychometric Aspects Description*

The psychometric aspects and their respective questions are as follows.

| **Psychometric Aspect** | **Psychometric Questions under this section** |
|---|---|
| Satisfaction with Life | ● In most ways, my life is close to my ideal<br>● The conditions in my life are excellent<br>● So far, I have got the important things I want in my life<br>● If I could live my life over, I would change almost nothing |
| Collectivism/ Allocentrism | ● Many people have directly or indirectly contributed to my progress in life<br>● I feel good when I co-operate with others<br>● When making a decision, I take other people's needs and feelings into account<br>● It is my duty to take care of my family, even when I have to sacrifice what I want<br>● Honesty is important to achieve success<br>● Social inequalities bother me |
| Individualism/ Idiocentrism | ● What I am today is solely because of my hard work and talent<br>● People should keep their troubles to themselves<br>● It is important that I do my job better than others<br>● Winning is everything<br>● I don't worry about others as long as I am happy<br>● It upsets me when my work is not recognized by others |
| Long-term Orientation | ● Traditional values are important for me<br>● I plan for the long-term<br>● I am willing to give up today's fun for future success<br>● I believe persistence is key to success<br>● Saving money is important to me |
| Short-term Orientation | ● It is okay to use shortcuts to get what you want<br>● I like to get quick results<br>● I would rather spend money today than save for future<br>● When my routine is disturbed, it upsets me |
| Materialism | ● I admire people who own expensive homes, cars, and clothes<br>● The things I own say a lot about how well I'm doing in life<br>● I aspire a luxurious and comfortable lifestyle<br>● My life would be better if I owned certain things I don't have<br>● I'd be happier if I could afford to buy more things |
| Spiritualism | ● I don't pay much attention to the material objects other people own<br>● I usually buy only the things I need<br>● I have all the things I really need to enjoy life |

| | |
|---|---|
| | ● My happiness does not depend on things I own<br>● There is a higher purpose to life than comfort and luxury<br>● I believe success in life does not mean becoming rich |
| Environmental Behaviour | ● I segregate waste before its disposal<br>● I try to conserve water<br>● I try to educate people I know about climate change<br>● I sign petitions related to environmental issues<br>● I try to conserve electricity<br>● I walk/cycle/use public transport to save fuel |

*Table 10: Psychometric Data Description*

**Appendix C: Codes and Documentation**

The source code for this project is attached as Jupyter Notebooks (PDF) in the submission. The following describes the files attached and their contents.

1. **Data Cleaning & EDA (Purchase)**
    - Used to perform data cleaning and exploration on the Purchase data
2. **Data Cleaning & EDA (Psychometric and Demographic)**
    - Used to perform data cleaning, exploration and preparation on the Psychometric and Demographic data
3. **Factor Analysis**
    - Used to perform factor analysis on the Psychometric and Demographic data
4. **Linear Regression - RFM Weights**
    - Used to generate a regression model to estimate RFM weights for Segmentation
5. **RFM Segmentation**
    - Used for Customer Segmentation using RFM and K-means, Transition Matrix Derivation and CLV calculation
6. **Classification Model**
    - Used to generate a classification model from descriptor variables
7. **Segment Differences**
    - Used to identify statistically significant segmentation and descriptor variables across segments defined