**Project Part One**

**Topic: Customer Behaviour Analysis**

**Department of Operation Management & Information System**

**OMIS 665: Big Data Analytics for Business**

**Professor: Dr. Ying Wang**

**Team   
Komal Chaudhari  
Vividha Chaudhari**

1. **Business Scenario**

Customer Behaviour Analysis is a detailed analysis of a company’s ideal customers. It helps a business to better understand its customers and makes it easier for them to modify products according to the specific needs, behaviors, and concerns of different types of customers.

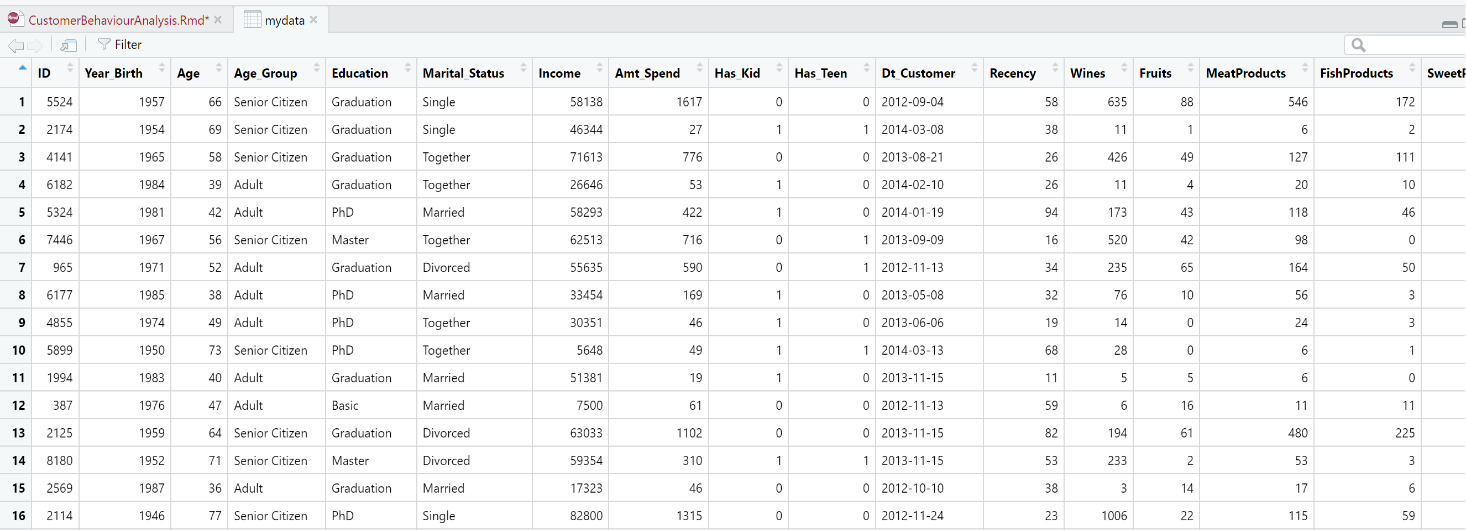
Customer Behaviour analysis helps a business to modify its product based on its target customers from different types of customer segments. For example, instead of spending money to market a new product to every customer in the company’s database, a company can analyze which customer segment is most likely to buy the product and then market the product only in that segment.

1. **About Data**

The dataset analyzed in this project provides a rich source of customer Behaviour information. The data set is a customer data set sourced from kaggle.com. The data contains 2240 observations and 33 variables.

A screen shot of a computer code

Description automatically generated



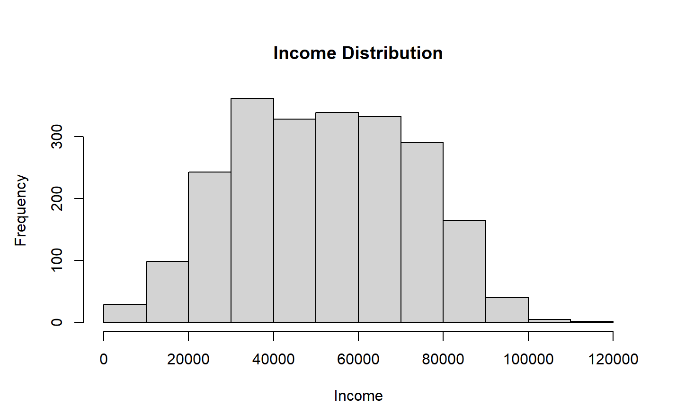
A screenshot of a computer

Description automatically generated

1. **Data Transformation**

In order to conduct a comprehensive analysis, the raw dataset underwent essential transformations. Notably, the "Income" variable contained numerous missing values. To address this, missing values were imputed using the median income value, chosen due to the presence of outliers, ensuring a robust estimation of missing income data. Additionally, leveraging the "Birth\_Year" column enabled the calculation of each customer's precise "Age," enriching the dataset with vital demographic information. These ages were further categorized into distinct groups: "Young Adult" (25-40 years), “Mid-Aged” (41-60) and "Senior Citizen" (61-100 years), offering insights into varied customer segments. The "Dt\_Customer" variable, signifying the customer's enrollment date, was harnessed to derive the "Enrollment\_Age," representing the age of customers at the time of their initial engagement with the company. Moreover, standardizing the date format (DD-MM-YYYY) in the "Dt\_Customer" column ensured consistency for seamless analysis. An important addition to the dataset was the introduction of the "Amt\_Spend" column, aggregating the total amount spent by each customer on products. This comprehensive spending metric was derived by summing values from key columns: "Wines," "Fruits," "MeatProducts," "FishProducts," "SweetProducts," and "GoldProds," providing a holistic view of customer purchasing behavior. These meticulous transformations not only enhanced the dataset's integrity but also laid the foundation for meaningful insights into customer behavior patterns and preferences.

1. **Descriptive Statistics**



Histogram of customer annual Income

The variable is normally distributed.

Mean income is $51620.

Min – Max income is $1730 - $113734.

A graph of a number of people

Description automatically generated

Histogram of customer Age

The variable is normally distributed.

Mean Age is 54

Min – Max Age is 27- 83.

A graph of age distribution

Description automatically generated

Histogram of customer enrollment age with company

The variable is normally distributed.

Mean Age is 44

Min – Max Age is 16 - 73.

A graph of a plot

Description automatically generated with medium confidence

Multimodal distribution

No skew

Outliers removed.

No evidence of dirty data

A graph of a function

Description automatically generated with medium confidence

Bimodal distribution

No skew

Outliers removed.

No evidence of dirty data

A graph of a number of points

Description automatically generated

Bimodal Distribution

Right skewed

Lognormal Distribution

No outliers

1. **Data Exploration**

**5.1 Understanding the Distribution of Customer Marital Status**

A graph of a bar chart

Description automatically generated with medium confidenceA graph of different colored squares

Description automatically generatedA colorful pie chart with text

Description automatically generated

The majority of the customer base is comprised of individuals who are married or in relationship, indicating that the products or services might be appealing to this demographic.

**5.2 Understanding the Distribution of Income vs Amount spend with different age groups**

The data suggests that the senior citizens earn more as well as they spend more than the mid-aged and young adults.

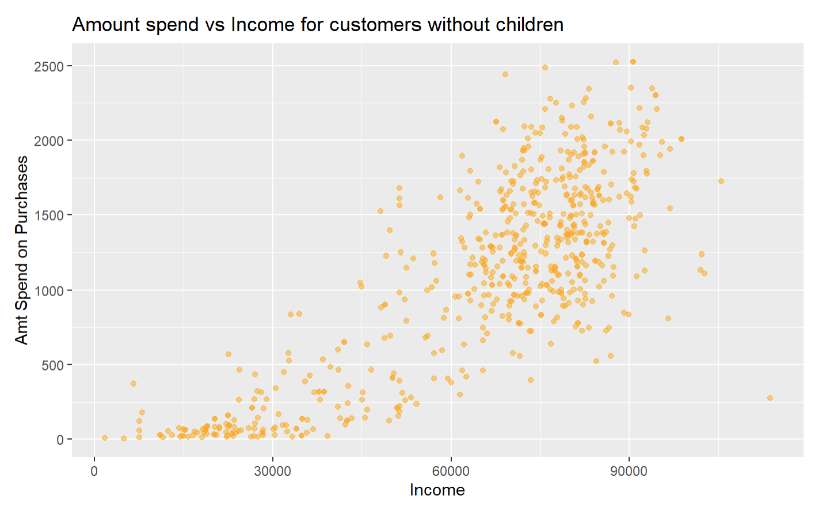
**5.3 Understanding customer responses to the campaigns**

A graph of blue bars

Description automatically generated with medium confidence

The data suggest that campaign 1, 3 ,4 and 5 has higher responses than campaign 2 indicating successful customer engagement. There is a further need to investigate what specific elements made the campaign successful.

Responses to campaigns

**A graph with orange dots

Description automatically generated5.4 Understanding the distribution of Income vs Amount spend for customers with and without children**

The above data suggests that the relationship between Amt\_Spend and Income remains the same whether you have any children or not.

* 1. **A screenshot of a computer

     Description automatically generatedLet’s predict the Amt\_Spend based on several predictor variables such as NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, and NumWebVisitsMonth**

**Conclusion:** We run a multiple regression model.This model suggests that the number of web and catalog purchases positively influences Amt\_Spend, while the number of deal purchases and web visits per month negatively influence it. The model overall explains a 70.31% of the variance in Amt\_Spend.

* 1. **A screenshot of a computer

     Description automatically generatedExamining the relationship between NumDealsPurchases and NumWebVisits**

**Conclusion:** We run a simple linear regression model. This model indicates a statistically significant positive relationship between the number of website visits (NumWebVisitsMonth) and the number of deals purchases (NumDealsPurchases). So we can say that if the number of website visits increases then the number of deal purchases also increases and vice versa.

* 1. **Examining the relationship between Amt\_Spend & Income and Amt\_Spend & Age**

**A close-up of a computer screen

Description automatically generatedA screen shot of a graph

Description automatically generatedCorrelation analysis between Income and Amt\_Spend:** There is a strong positive correlation between Income and Amt\_Spend with a correlation coefficient of 0.8202216. We can say that if one variable increases, the other variable tends to increase as well, and vice versa. The value of coefficient of correlation, 0.8202216 being close to 1 suggests a very strong positive linear relationship between the variables.

**A graph of red dots

Description automatically generatedCorrelation analysis between Age and Amt\_Spend:** There is a weak positive correlation between Age and Amt\_Spend with a correlation coefficient of 0.1160903. We can say that if one variable increases, the other variable tends to increase slightly, but the relationship is extremely weak. We can also say that the correlation coefficient is close to zero which indicates a lack of substantial linear relationship between the variables. In conclusion we can say that younger and older customers both contribute to sales, but the correlation is weak, indicating age is not a significant factor in determining spending patterns.

1. **Hypothesis**

**Test 1:** We want to know if there is a significant difference in the amount spent on wines based on marital status.

Null Hypothesis (H0): There is no significant difference in the mean amount spent on wines among different marital status categories.

A screen shot of a computer

Description automatically generatedAlternative Hypothesis (H1): There is a significant difference in the mean amount spent on wines among different marital status categories.

**Conclusion**: The result contains p-value (0.368) for degrees of freedom 7. As this p-value is greater than the alpha (0.05), therefore we fail to reject the null hypothesis and conclude that there is no significant difference in the mean amount spent on wines among different marital status categories. We can say that marital status does not significantly influence the amount spent on wine.

**Test 2:** We want to know if age, income, marital status, has kids or teenagers, education have any impact on Amt\_Spend.

Null Hypothesis (H0): The coefficient for predictor variables is equal to zero, indicating no effect on Amt\_Spend

Alternative Hypothesis (H1): The coefficient for atleast one predictor variable is not equal to zero, indicating a significant effect on Amt\_Spend.

A screenshot of a computer

Description automatically generated

**Conclusion**: The regression model suggests that Income, having children (Has\_Kid), having teenagers (Has\_Teen), and having Basic education significantly (p-value <0.05) influence the amount spent (Amt\_Spend). Other variables, including Age, Marital\_Status, and different education levels, do not show a significant impact.

**Test 3**: We wanted to test if the income was different for different Age\_Groups

Null Hypothesis (H0): The two means are same.

A computer screen shot of a computer code

Description automatically generatedAlternative Hypothesis (H1): The two means are different.

A graph with a row of rectangles

Description automatically generatedA graph of a graph

Description automatically generated

**Conclusion**: We conducted a t-test with categories “Mid-Aged” and “Senior Citizen” as they were normally distributed and filtered category “Young Adult” as it was not normally distributed. As a result of the t-test we could conclude that the two means are different. We can say that the Senior Citizens earn more than the Mid-Aged.

1. **Conclusion**

After performing descriptive analysis through various visualizations and performing statistical test on different variables, here are our findings:

* Understanding the predominant marital status can help in deciding the marketing strategies. For instance, we can have campaigns focusing on family-oriented products that might resonate well with married or partnered individuals.
* Tailoring marketing strategies to appeal to the senior citizens and high-income individuals’ income can exhibit substantial spending habits.
* Customers tend to spend more when making web and catalog purchases, while they limit spending during deal purchases and frequent web visits.
* Regardless of marital status, customers show similar spending patterns in this category.
* Spending patterns are notably affected by having children and basic education. These factors significantly influence expenditure. However, age, marital status, and varying education levels do not exert a substantial impact on customer spending behavior.