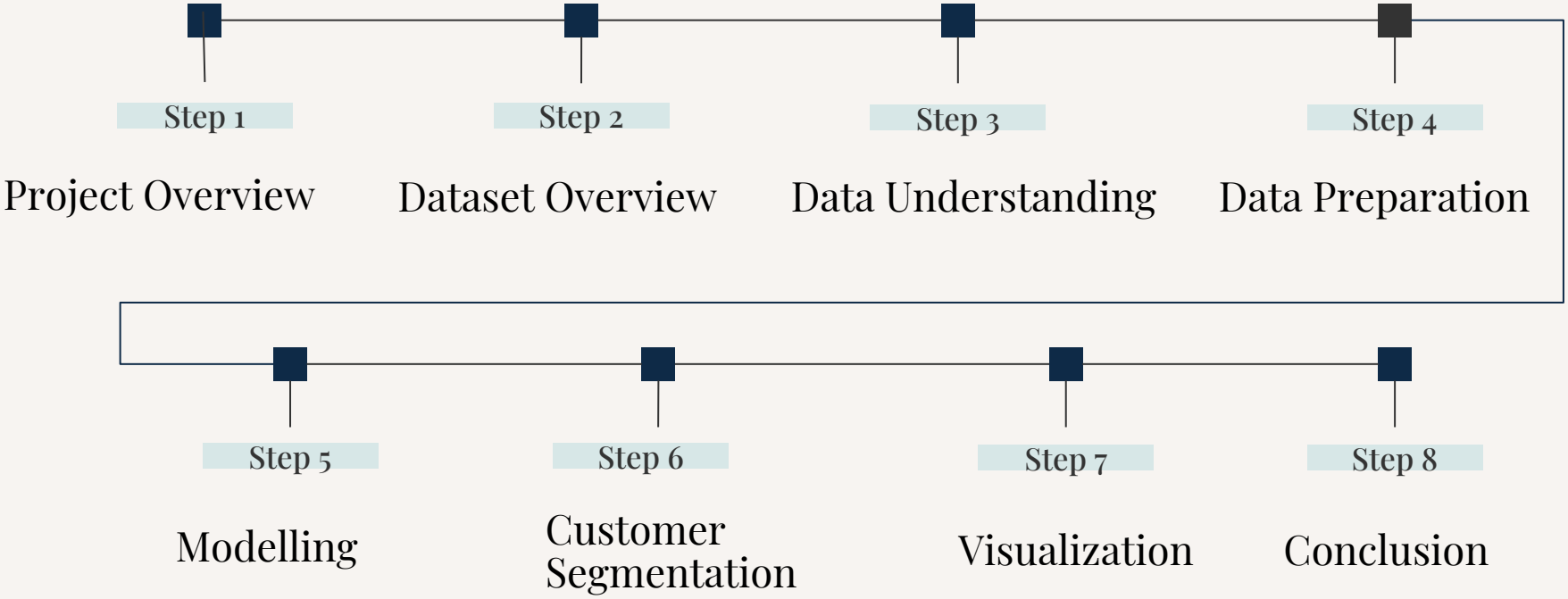
Customer Data Analysis and Predictive Modelling

Conducted by

Nazila Fazeli : M.Sc. Artificial Intelligence

Our process


Step 1
Project Overview
Step 2
Dataset Overview
Step 3
Data Understanding
Step 4
Data Preparation
Step 5
Modelling
Step 6
Customer Segmentation
Step 7
Visualization
Step 8
Conclusion

GOALS

1. **Predict Total Spending:** Build a model to predict customer spending.
2. **Predict Campaign Responses:** Identify customers likely to respond positively to marketing campaigns.
3. **Customer Segmentation:** Divide customers into meaningful clusters based on behaviours.
4. **Visualization:** Develop Dashboards to visualize insights and support strategic decisions.

Dataset Overview

- Data Summary:
  - Name: Customer Personality Analysis
  - Rows: 2,240 ↠ 1,946 (after cleaning).
  - Variables: 27.
- Dataset Categories:

01

Demographic Variables:

- ID: Unique customer identifier.
- Year_Birth: Year of birth.
- Education: Education level.
- Marital_Status: Marital status.
- Income: Yearly income.
- Kidhome: Number of children.
- Teenhome: Number of teenagers.
- Dt_Customer: Date of enrollment in the program.
- Recency: Days since last purchase.

02

Purchase Behavior Variables:

- Spending on product categories: MntWines, MntFruits, MntMeatProducts, etc.

03

Marketing Interaction Variables:

- Responses to campaigns: AcceptedCmp1 to AcceptedCmp5, Response.

04

Channel Engagement Variables:

- Purchases through channels: NumWebPurchases, NumCatalogPurchases, NumStorePurchases.
- Visits: NumWebVisitsMonth.

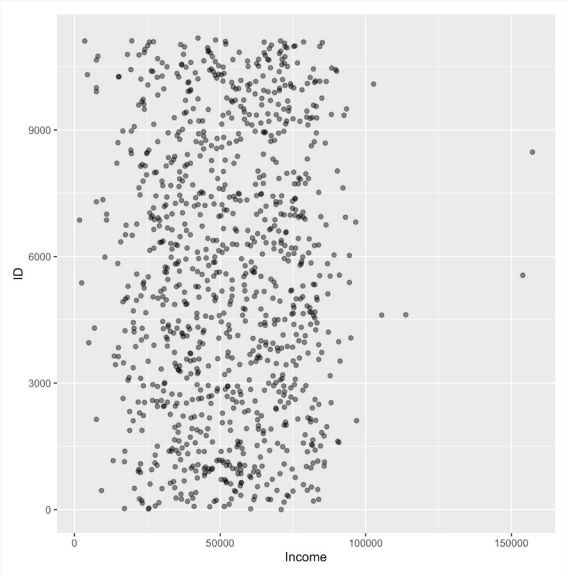
05

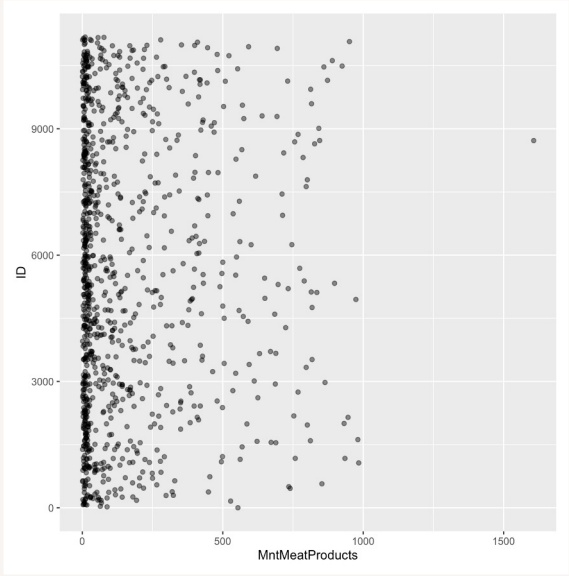
Other Variables:

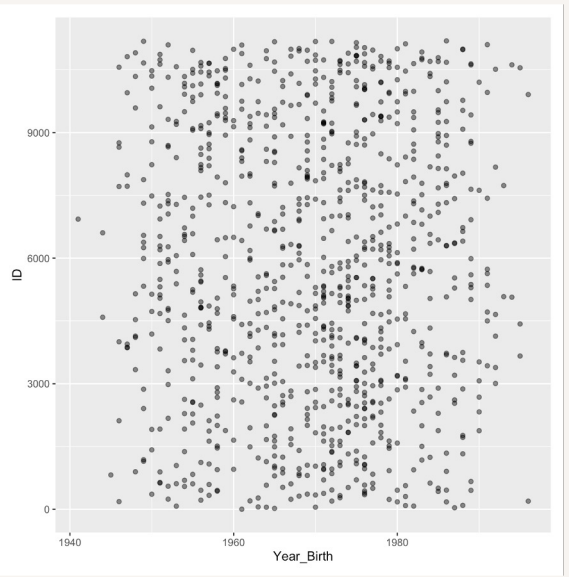
- Complain: Customer complaints.

Data Cleaning

1. Removed rows with missing Income.
2. Identified and removed outliers (e.g., invalid Year_Birth, extreme Income).
3. Applied Z-Score and IQR for final outlier removal, choosing Z-Score for better retention of data.





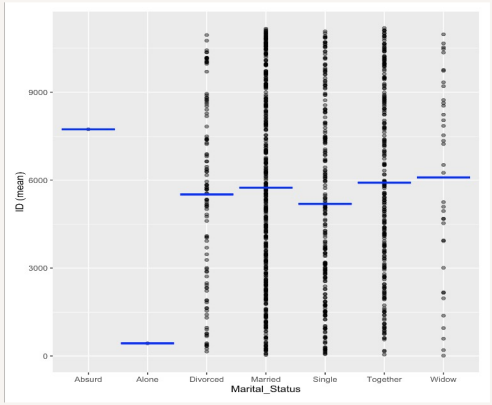


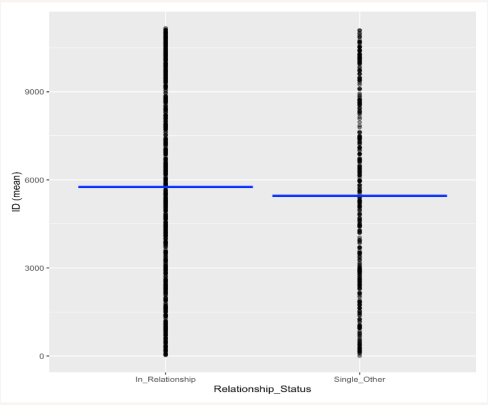
Data Preparation

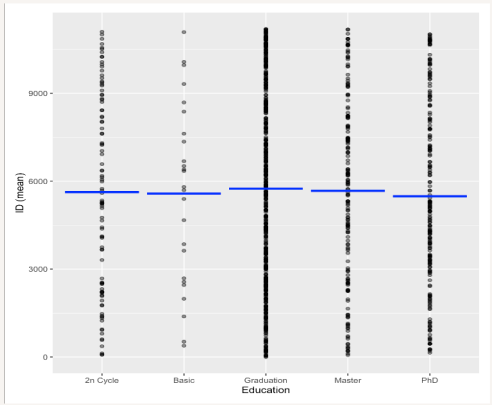
Feature Engineering

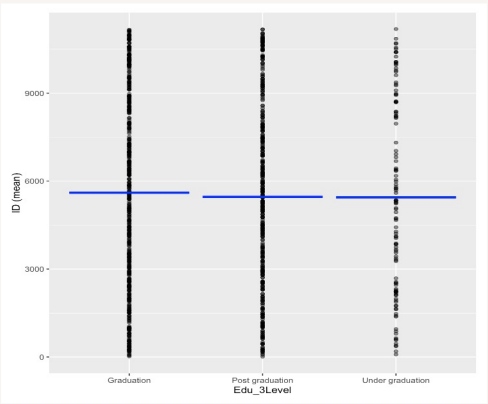
Created new variables:

- Age (from Year_Birth)
- Total Spend
- Family Size
- Total Accepted Campaigns
- Education Level
- Relationship Status





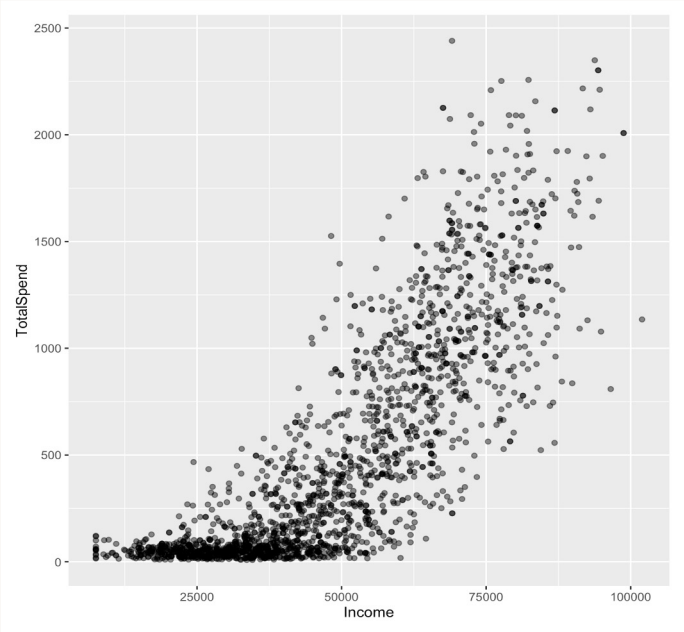


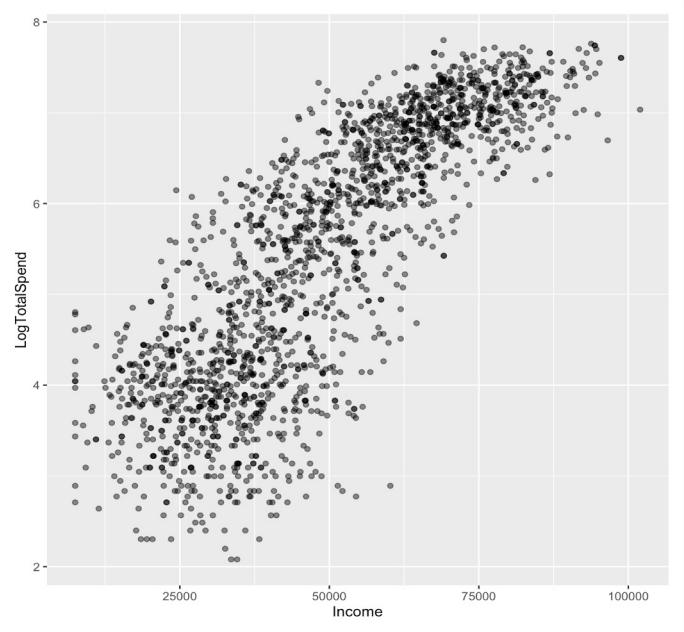


Linear Regression: Objective

Predicting Total Spending

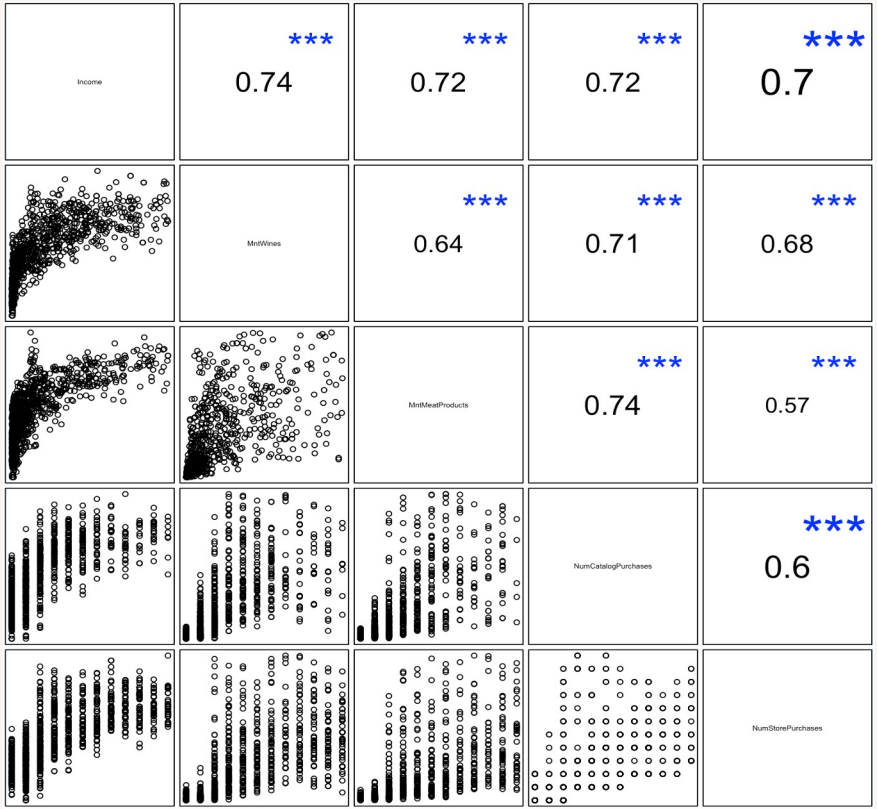
Used Log Total Spend for better accuracy.





Linear Regression: Correlation Analysis

Analysed correlations among independent variables, confirmed that higher correlations did not impact model accuracy, allowing inclusion of correlated variables.



Linear Regression: Feature Selection

Used Stepwise selection and trial-and-error to find the best combination of independent variables with high R2. Settled on nine variables for optimal performance.

Number of Variables	Explanatory Variables	R^2 Model
8	Income + Kidhome + Teenhome + MntWines + MntGoldProds + NumDealsPurchases + NumWebPurchases + NumStorePurchases	91%
9	Income + Kidhome + Teenhome + MntWines + MntMeatProducts + MntGoldProds+ NumDealsPurchases + NumWebPurchases + NumStorePurchases	91.6%
10	Income + Kidhome + Teenhome + MntWines + MntMeatProducts + MntGoldProds+ NumDealsPurchases + NumWebPurchases + NumCatalogPurchases + NumStorePurchases	91.8%
12	Income + Kidhome + Teenhome + Dt_Customer + MntWines + MntMeatProducts+ MntFishProducts + MntGoldProds + NumDealsPurchases + NumWebPurchases + NumCatalogPurchases + NumStorePurchases	92.1%

Linear Regression: Evaluation

Displayed R2 values and performance results for Train and Test data, showing high accuracy (above 90%) in predicting **Total Spend**

Number of Variables	Explanatory Variables	R^2 Model	R^2 Train Model	R^2 Test
9	Income + Kidhome + Teenhome + MntWines + MntMeatProducts + MntGoldProds + NumDealsPurchases + NumWebPurchases + NumStorePurchases	91.6%	91.7%	90.9%

Logistic Regression

Attempted to model customer complaints but found insufficient data.

Created a model to predict Response to the latest campaign with good Sensitivity and Specificity.(Using 10 Variables)

Threshold > 0.09

Show

	ResponsePred		
Response	**X0**	**X1**	**Total**
	All	All	All
0	312	117	**429**
1	4	64	**68**
Total	**316**	**181**	**497**

Thereshold > 0.09 Baseline & Accuracy

Show

	ResponsePred		
Response	**X0**	**X1**	**Total**
	All	All	All
0	0.628	0.235	**0.863**
1	0.008	0.129	**0.137**
Total	**0.636**	**0.364**	**1.000**

Thereshold > 0.09 Specificity & Sensitivity

Show

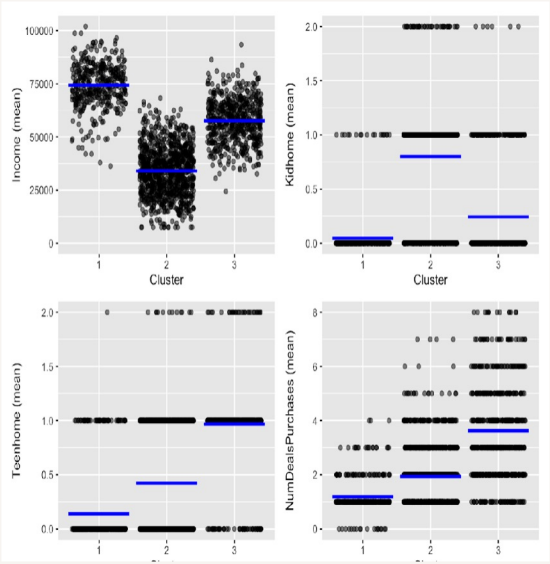
	ResponsePred		
Response	**X0**	**X1**	**Total**
	All	All	All
0	0.727	0.273	**1.000**
1	0.059	0.941	**1.000**
Total	**0.636**	**0.364**	**1.000**

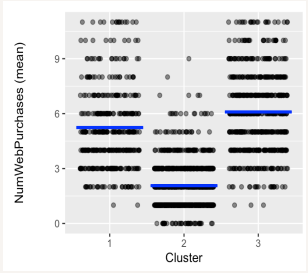
Customer Segmentation

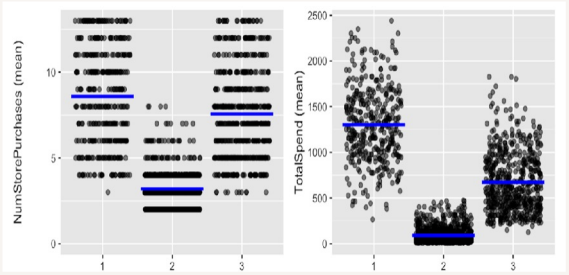
High Spenders: Loyal and consistent customers.

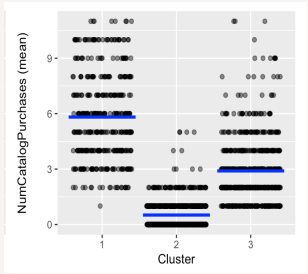
Low Spenders: Minimal activity or churn risks.

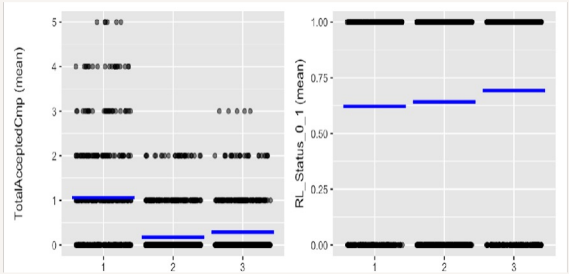
Moderate Spenders: Average engagement.

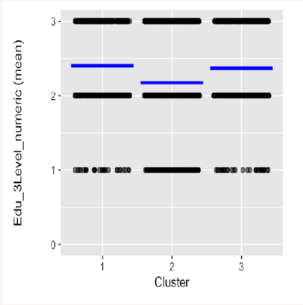












Data Visualization in Tableau

- Created various sheets for demographic and behavioral analysis:

Age, Income, Total Spend, Product Purchases, Marital Status, etc., to explore customer trends.

Conclusion

Main Findings:

- **High-income customers** are the main drivers of total spend across product categories.

- **Linear Regression** effectively predicted total spending with high accuracy (**R^2 > 90%**).

- **Logistic Regression:**

    - Results for predicting campaign responses were **not optimal**, as the model struggled with limited variability in campaign responses.
    - Further tuning or additional data may improve its performance.

- **Clustering:**
  Successfully identified **three customer segments**, providing actionable insights for targeted marketing.