**Customer Segmentation and Targeted Marketing Analysis**

**Data Collection**

This project selects a dataset focusing on customer behaviour and customer responses to marketing campaigns. The variables that it does have to contain are their demographic details (age, income, education, marital status), spending behaviour (amount spent on particular product categories), and reaction to marketing campaigns (accept or reject an offer). The dataset was chosen because it’s relevant to real-life business situations, especially in retail, e-commerce and the telecommunication industry, where Targeting and segment searches are essential. Businesses can use this data to see who their customers are in distinct ways, which will effectively target marketing efforts for them. Customer segmentation by this type is essential to effective customer engagement, as well as to increase sales and enable marketing strategies to be focused on specific groups.

**Data Exploration and Preprocessing**

Data exploration and preprocessing start with getting familiar with the dataset. These include inspecting the data structure to see what types the data contains, like categorical, numerical or DateTime, and understanding the feature. For example, the dataset includes numerical features like income, the frequency of purchases, the amount spent in different product categories, and categorical features like education level and marital status. The next thing is to deal with missing values. For this case, the income column has missing values that can be imputed with the mean or median not to drop rows that may have valuable values.

Furthermore, Education and Marital\_Status are encoded as categorical variables using one hot encoding. This converts these categories into binary columns, making them easy to analyze using machine learning algorithms. Feature engineering is an essential part of preprocessing. For instance, the Dt\_Customer column that denotes when the customer became a company member is changed into a new feature, Customer\_Age, the number of months since registration. To ensure that no single feature dominates the analysis given K-means for an example of a distance-based model, all numerical features are standardized (scaled to mean 0 and standard deviation 1).

**Task Selection**

After preprocessing the data, the next step is to decide what analytical task(s) will deliver the most significant value. We have chosen the main task from clustering to analysis for this case. This allows us to cluster customers according to their buying behaviour, demographic data, and responses to previous marketing campaigns. Using the K-means algorithm, we obtain distinct customer segments. Businesses can develop more personalized marketing strategies based on how customers are grouped with similar characteristics, leading to a higher engagement and conversion rate. Another possible task we could study is classification, where we could build a model to predict whether a customer will respond to a future marketing campaign. We could use customer features like income, age, etc. and analyze how they have responded to our past campaigns. Nevertheless, clustering is first selected because it directly addresses the segmentation problem, the main focus of the dataset. After groups of customers are identified, predictive modelling could be used to target campaigns appropriately.

**Methodology**

This project's methodology begins with feature selection, where we chose pertinent variables within the dataset, which would assist us in segmenting customers effectively. Next, the K-means clustering algorithm is applied to a combination of demographics and spending behaviour features. The K mean application's first part is to find the number of clusters. It is typically done through the Elbow Method, where we plot inertia (the sum of squared distances of samples to their closest cluster centre) over different clusters. The point where the 'elbow' in the plot occurs denotes the best number of clusters. Also, we use another high-quality metric to evaluate each clustering, i.e., the Silhouette Score. If the silhouette score is high, the cluster is well-defined, and each customer should fall into a particular group. After determining the number of clusters, we run the K-means algorithm with the resulting clusters and then analyse them to discover the characteristics of each segment. For example, the first cluster contains high-income customers who buy many luxury things, and the second cluster contains budget-conscious customers who randomly purchase these things. These insights can then be leveraged to create targeted marketing initiatives towards each customer group.

**Reference**

<https://www.kaggle.com/code/karnikakapoor/customer-segmentation-clustering/input>