Problem Statement: Findings relationships among the descriptive attributes in account details and labels datasets.

Overview:

1. Read and understand the data
2. Clean the data
3. Prepare the data
4. Modelling
5. Final Analysis

Read and understand the data:

1. Import all necessary libraries
2. Use functions to understand the data

Clean the data:

1. Find the null values in each column. Dropped the columns whose missing value percentage is greater than 50%.
2. Plot the graphs to visualize the data
3. Check the correlation between columns in the datasets.

Prepare the data:

1. Consider RFM method for clustering to understand the behaviour of different patient IDs. R (Recency),F (Frequency),M (Monetary)
2. Consider 3 numerical columns **Totpayment, Amount\_Due,** **AR\_Percentage for KMeans Clustering.**
3. Done Outlier treatment and rescaling

Modelling:

1. Consider n\_clusters= 4 for reference.
2. To find the optimal number of clusters I used elbow curve method.
3. To cross check cluster number I also used silhouette analysis method
4. Assign the label and plot the graphs

Final Analysis:

1. Amount\_Due, TotAmount are more in cluster no:1 compared to other clusters.
2. AR\_Exist and AR\_Dollar\_threshold are most correlated.
3. TotPayment and Expect\_amount are most correlated.