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 column of missing value percentage which 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. Assign the label and plot the box plots to identify the cluster ranges.

Final Analysis:

1. In account\_details\_labels data set in Type column consists of data that Type 2 (62.55%) count is more than Type 1 (37.45%)
2. Code MR080 has highest value of Amount\_Due.
3. Top 10 Amount\_Dues against Code is 91% of the overall Amount\_Due.
4. MR080 Code has highest value of AR\_Dollar\_threshold.
5. Highest AR\_Dollar\_threshold value occupies 48% of the overall AR\_Dollar\_threshold.
6. Amount\_Due & TotAmount values are more in cluster no:1 compared with Cluster 0 and cluster 2.