# Project 3

This project includes but not limited to the following:

We start by the EDA part:

1. Check Duplicates
2. Check Nulls
3. We discover that CustomerID is just an ID, but before deleting it, we want to make sure that it has n unique values since a replicate CustomerID indicates something is wrong
   * Since it has no duplicates, we drop it
4. We map the Gender into 0 & 1 [0🡪Male] [1🡪Female]
5. Plot pair plot which helps:
   * To get an idea of the number of cluster
   * To get an idea of the linear correlation between the features
6. Plot box plot to investigate outliers
   * We discover that there is just one outlier so we did not take an action
7. Heat map to see Pearson’s correlation between the features
8. Plot bigger scatter plots
9. The conclusion of the EDA are:
   * - It seems there are no high correlations between the variables
   * - Scaling seems unnecessary but better since clustering depend on distances
   * - Based on the scatter plot: customers can be segmented into 2-5 clusters
10. Create a function of Hopkins test which can be used for tendency of clustering, where >0.5 is support that idea
    * We use the whole dataset for the test since its small [n=200]
11. To determine the number of clusters: Elbow method of inertia & elbow method of distortion & Silhouette’s test were performed By using the tree methods into account, 5 clusters seems the better number of clusters [for the two pair of features]
12. We plot the clusters using two features
13. Then, HC has been done, dendrogram of the four available methods:
    * Ward
    * Complete
    * Average
    * Single
14. Ward Method with three cluster was used
15. Conclusion based on the kmean with 5 clusters has been written